

DOCUMENT RESUME

ED 121 805

TM 005 232

AUTHOR Kalisch, Stanley James, Jr.
TITLE The Comparison of Two Tailored Testing Models and the Effects of the Models' Variables on Actual Loss.
PUB DATE Dec 74
NOTE 150p.; Ph.D. Dissertation, Florida State University

EDRS PRICE MF-\$0.83 HC-\$7.35 Plus Postage
DESCRIPTORS Ability; Branching; *Comparative Analysis; *Computer Oriented Programs; Decision Making; Efficiency; Instruction; *Models; Prediction; Probability; *Response Style (Tests); Simulation; Statistical Analysis; *Testing
IDENTIFIERS *Tailored Testing

ABSTRACT

The four purposes of this study were: (1) To compare two versions of a tailored testing model similar to one suggested by Kalisch (1974); (2) To identify levels of the variables within the two versions, which produce an efficient tailored testing procedures; (3) To compare, within each version, the results obtained when employing relatively small data bases with those obtained using large data bases; and (4) To determine, within each version, whether any of the tailored testing procedures are more efficient than obtaining the responses to all the items from each examinee. The models specified procedures by which the correctness of an examinee's responses to items of a test are predicted by requiring an examinee to answer as few test items as possible. The study employed computer simulations, using existing data collected from examinees who had answered all the test items. It was concluded that the model, and the levels of variables within the models, produce differential efficiency, dependent on the evaluation matrix. The testing procedures employing small data bases are not necessarily as efficient as the procedures employing large data bases, but for the evaluation matrices used in the study, the tailored testing procedures were more efficient than obtaining an examinee's response to every item. (RC)

* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *

THE COMPARISON OF TWO TAILORED TESTING MODELS AND THE
EFFECTS OF THE MODELS' VARIABLES ON ACTUAL LOSS

BY

STANLEY JAMES KALISCH, JR.

Computer Based Education Company
Control Data Corporation

A Dissertation submitted to the Division of
Educational Management Systems, Program of
Educational Evaluation and Research Design,
at Florida State University - December, 1974.

Copyright © 1974 by Stanley James Kalisch, Jr.
All rights reserved.

PERMISSION TO REPRODUCE THIS COPY-
RIGHTED MATERIAL HAS BEEN GRANTED BY

Stanley Kalisch Jr.
TO ERIC AND ORGANIZATIONS OPERATING
UNDER AGREEMENTS WITH THE NATIONAL IN-
STITUTE OF EDUCATION. FURTHER REPRO-
DUCTION OUTSIDE THE ERIC SYSTEM RE-
QUIRES PERMISSION OF THE COPYRIGHT
OWNER.

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRO-
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIGIN-
ATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRESENT
OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY.

Dedicated to my wife, Cindy--
without whom this dissertation would have never been written.

THE COMPARISON OF TWO TAILORED TESTING MODELS
AND THE EFFECTS OF THE MODELS'
VARIABLES ON ACTUAL LOSS

(Publication No.)

Stanley James Kalisch, Jr., Ph.D.
The Florida State University, 1974

Major Professor: Howard W. Stoker

Two tailored testing models, intended for use in instructional testing, were developed and investigated. The models specified procedures by which the correctness of an examinee's responses to items of a test are predicted by requiring an examinee to answer as few test items as possible. The study employed computer simulations, using existing data collected from examinees who had answered all the test items.

The models employed item conditional difficulties, beta distributions, loss functions, and decision theory. The models required data bases consisting of S_s ' responses to each item in order to determine conditional difficulties and to compute expected losses for each decision to be made for unrepresented items. These decisions were (1) predict the item response to be correct, (2) predict the response to be wrong, and (3) obtain a response. If the expected loss for a prediction was less than that for obtaining a response, the prediction was noted.

After all predictions were noted, the item with a conditional

difficulty closest to 0.5 was presented to the examinee. The procedure continued until each item response was obtained or predicted. Unlike Model 1, Model 2 required reconsideration of each prediction after each response by an examinee.

The models were compared for efficiency, the total actual loss per examinee. Actual losses were computed from values in an evaluation matrix. Three different independently obtained evaluation matrices were used. Different evaluation matrices applied to the same results of the tailored testing procedures produced different conclusions. In relation to the matrices, the models proved to be equally efficient except in one case in which Model 1 was more efficient than Model 2.

An additional purpose of the study was to investigate the effects of variables in the testing models on efficiency. The presentation of one or as many as five items prior to attempting prediction produced no differential efficiency related to each evaluation matrix, except one in which the presentation of only one item was most efficient.

Another variable investigated was critical conditional difficulty. In order to implement the testing procedures, it would be necessary for the decision makers to preset the critical conditional difficulties. The levels of the critical conditional difficulties considered were .80, .85, .90, and .95. The conclusions were greatly dependent upon the evaluation matrices. With respect to the first matrix, no differential effects occurred. In relation to the second matrix, the lowest level produced the

most efficient procedure. With the third matrix, the results were different for the two models. In Model 1 the lowest level produced the most inefficient procedure; the other levels were equally efficient. In Model 2 the third level provided the most efficient procedure.

The efficiency of the testing procedures employing small data bases of 100 or 200 was equal to the efficiency utilizing large data bases of 2000 for only one matrix. For the other matrices, the smaller data bases reduced efficiency. For each evaluation matrix, the testing procedures employing data base sizes of 100 or 200 were more efficient than obtaining responses to all items from an examinee.

On the basis of the study's results, it was concluded that the model, and the levels of the variables within the models, produce differential efficiency, dependent upon the evaluation matrix. The testing procedures employing small data bases are not necessarily as efficient as the procedures employing large data bases, but for the evaluation matrices used in the study, the tailored testing procedures were more efficient than obtaining an examinee's response to every item.

ACKNOWLEDGEMENTS

The completion of a dissertation is the culmination of years of work. It is also a time when one reflects upon those years and upon the many individuals who have contributed to the author's fulfillment of a dream. A dissertation is not the product of one individual, but that of numerous individuals, many of whom, probably unknowingly, have very significantly influenced the author and whose influence is reflected in this dissertation.

The author is extremely indebted to Dr. Howard Stoker, who in directing this dissertation, provided the author with much needed advice and encouragement. Dr. Stoker's genuine interest and humanistic qualities toward the author and his work have been greatly appreciated and have engendered a feeling of lasting gratitude.

Each of the author's committee members has been an asset in producing the final product. Dr. Jacob Beard provided comments and suggestions that have greatly influenced the direction of the author's work. The author's assistantship for three years with Dr. Paul Merrill was most beneficial and enriching. Dr. E. Thomas Denmark assisted in ascertaining the correctness of the mathematical aspects of the study.

The author is also appreciative of the positive reaction and encouragement provided by Dr. John Hills, in whose class the author first began the development of his tailored testing model.

A very special expression of thanks is due to Robert Jordan, who suggested a means of simplifying some of the formulas used in the study. Without obtaining the simplified forms, the study may have never been completed.

The author is most appreciative of the efforts of Dr. Gambir Shrestha in providing the data used in the study. Dr. Shrestha gave much of his time in providing the author with the data and information related to the testing program in which the data was obtained.

Drs. Walter Dick and Nelson Towle provided evaluation loss matrices used in the study. Their cooperation was most greatly appreciated.

The author is indebted to the staff of the Center for Educational Design, particularly to Drs. John Merrill, Paul Merrill, and Nelson Towle, and to Mrs. Betty Wright for their support, understanding, and friendship during his work at Florida State University.

The greatest expression of my appreciation is directed to my wife, Cindy. Her assistance in entering computer program code and the typing of the dissertation was extremely helpful. Her encouragement and compassion provided me the incentive to continue. Her

resolve in my completing this dissertation has brought us to
this moment in sharing in the fulfillment of a long-time dream
of ours.

TABLE OF CONTENTS

ABSTRACT	iii
ACKNOWLEDGEMENTS	vi
LIST OF TABLES	xi
LIST OF FIGURES	xiii
Chapter	
I. INTRODUCTION	1
II. REVIEW OF RELATED LITERATURE	4
Purposes of Tailored Testing	4
Assumptions	6
Branching Strategies	7
Scoring	9
Sizes of the Measurement Sections	10
Kalisch's Tailored Testing Model	11
III. THE TAILORED TESTING MODELS	15
Conditional Difficulty of an Item	15
Item Conditional Difficulty and the Beta Distribution	18
Loss Functions	21
A Linear Loss Assumption	26
A Non-Linear Loss Assumption	26
Using an Evaluation Loss Matrix for Decision Making	28
Determining Expected Loss	30
Selecting Loss Values for the Decision Loss Matrix	32
Model 1	38
Model 2	42
IV. THE PROBLEM	45
V. METHODS AND PROCEDURES	47
Dependent Variables	49
Independent Variables	51
The models	51
The critical conditional difficulty values	51

Number of items to be presented prior to predictions	54
Number of subjects for the data base	54
Part 1 of the Study	54
Part 2 of the Study	57
VI. RESULTS	62
Part 1 of the Study	62
Part 2 of the Study	84
VII. DISCUSSION	91
Which of the Two Tailored Testing Models is More Efficient?	93
How Many Items Should be Presented Before Attempting Prediction?	93
What Critical Conditional Difficulty Values Should be Selected?	94
Model 1	94
Model 2	95
Does a Data Base Consisting of Fewer than 1000 Sets of Item Responses Provide as Efficient a Procedure as a Data Base Consisting of 2000 Sets of Responses?	95
Are the Tailored Testing Procedures, Using Data Bases Consisting of Fewer than 1000 Sets of Item Responses, More Efficient than Presenting All the Items to an Examinee?	96
Implications of the Study	96
Efficiency of the Testing Procedures	97
Eliminating the Decision Loss Functions	99
Applicability of the Models to Instructional Testing	100
VIII. SUMMARY AND CONCLUSIONS	102
APPENDICES	105
A. REDUCTION OF FORMULA 12 TO FORMULA 13	106
B. REDUCTION OF FORMULA 16 TO FORMULA 17	110
C. DIRECTIONS GIVEN TO INDIVIDUALS SUPPLYING EVALUATION MATRICES	115
D. A TRANSFORMATION TECHNIQUE OF THE INDEPENDENT VARIABLES OF A REGRESSION EQUATION	122
E. DESCRIPTIVE STATISTICS FOR FACTORS AND INTERACTION IN THE ANALYSES OF VARIANCE OF PART 1 OF THE STUDY	127
REFERENCES	130
VITA	133

LIST OF TABLES

1. Subject by Item Response Matrix for Ten Subjects and Three Items	17
2. Matrix of Loss Values for Items with Conditional Difficulty Values of 1 or 0	22
3. Matrix of Loss Values for Determining Actual Loss	24
4. Values of Factors for Formulas 21 and 22	35
5. Expected Loss Values for Obtaining a Response for $L_p = 183$, $L_0 = 100$, and Various Data Base Sizes	39
6. Adjusted Critical Conditional Difficulty Values for Various Data Base Sizes	40
7. Evaluation Loss Matrix A	50
8. Evaluation Loss Matrix B	50
9. Evaluation Loss Matrix C	51
10. Critical Loss Values for Prediction for Selected Critical Conditional Difficulties	52
11. Decision Loss Matrix Used for Critical Conditional Difficulty Level 1	53
12. Descriptive Statistics of Total Actual Loss for Model 1 in Part 1 of Study	63
13. Descriptive Statistics of Total Actual Loss for Model 2 in Part 1 of Study	64
14. Descriptive Statistics of the Results for Model 1 in Part 1 of the Study	65
15. Descriptive Statistics of the Results for Model 2 in Part 1 of the Study	66
16. Analysis of Variance for Total Actual Loss Related to Evaluation Loss Matrix A	68
17. Analysis of Variance for Total Actual Loss Related to Evaluation Loss Matrix B	69
18. Analysis of Variance for Total Actual Loss Related to Evaluation Loss Matrix C	70
19. Results of the Newman-Keuls Test for Total Loss for Critical Conditional Difficulty Levels Within Model 1 and Model 2 Relative to Evaluation Loss Matrix A	71
20. Results of the Newman-Keuls Test for Total Loss for Critical Conditional Difficulty Levels Within Model 1 and Model 2 Relative to Evaluation Loss Matrix B	72
21. Results of the Newman-Keuls Test for Total Loss for Number of Items Presented Prior to Prediction Relative to Evaluation Loss Matrix B	73

22.	ANOVA Table for Significance Tests for Coefficients and for Lack of Fit for Assumed Linear Model for Evaluation Loss Matrix B and Model 1	75
23.	ANOVA Table for Significance Tests for Coefficients and for Lack of Fit for Assumed Linear Model for Evaluation Loss Matrix B and Model 2	76
24.	Results of Regression After Transformation of the Independent Variables	78
25.	ANOVA Table for Significance Tests for Coefficients and for Lack of Fit for Regression Equation Obtained from Iterative Procedure for Evaluation Loss Matrix A Within Model 1	79
26.	ANOVA Table for Significance Tests for Coefficients and for Lack of Fit for Regression Equation Obtained from Iterative Procedure for Evaluation Loss Matrix B Within Model 1	80
27.	ANOVA Table for Significance Tests for Coefficients and for Lack of Fit for Regression Equation Obtained from Iterative Procedure for Evaluation Loss Matrix C Within Model 1	81
28.	ANOVA Table for Significance Tests for Coefficients and for Lack of Fit for Regression Equation Obtained from Iterative Procedure for Evaluation Loss Matrix A Within Model 2	82
29.	ANOVA Table for Significance Tests for Coefficients and for Lack of Fit for Regression Equation Obtained from Iterative Procedure for Evaluation Loss Matrix C Within Model 2	83
30.	Point Estimates of Total Actual Loss for Conditional Difficulty by Item Levels Within Each Model	85
31.	Descriptive Statistics of Total Actual Loss Involving Data Bases of Different Sizes	86
32.	Comparison of Actual Losses for Simulations Employing Data Bases of Size e with Losses Obtained by Using a Data Base of Size 2000	87
33.	Results of One-Tailed t-tests to Compare Total Actual Loss Obtained for Data Bases of Size 100 and Loss Incurred for Presenting All Items	89
34.	ANOVA Summary for Transformation Technique	126
35.	Descriptive Statistics for Total Actual Loss Relative to Evaluation Loss Matrices A, B, and C for the Three Analysis of Variance Factors	128
36.	Descriptive Statistics for Total Actual Loss Relative to Evaluation Loss Matrices A, B, and C for the Model by Item Interaction Term for the Analysis of Variance	129

LIST OF FIGURES

1. Three beta distributions	20
2. Two linear loss functions	27
3. A non-linear loss function	28
4. The three loss functions used in the study	36

CHAPTER I

INTRODUCTION

The four purposes of the study were:

1. To compare two versions of a tailored testing model similar to one suggested by Kalisch (1974).
2. To identify levels of the variables within the two versions, which produce an efficient tailored testing procedure.
3. To compare, within each version, the results obtained when employing relatively small data bases with those obtained using large data bases.
4. To determine, within each version, whether any of the tailored testing procedures are more efficient than obtaining the responses to all the items from each examinee.

Kalisch (1974) contended that the tailored testing model he proposed may prove to be simpler than the procedures employed by Ferguson (1970), Lord (1970, 1971), Waters and Bayroff (1971). Kalisch's model may also be applicable to instructional testing. The model specifies a procedure by which the correctness of an examinee's responses to items of a test are predicted by presenting as few items as possible to the examinee. A validation study reported by Kalisch (1974) required, on the average, 5s to respond

to eight items, and correctly predicted the correctness of responses to 10 of the remaining 12 items of the test. The results were probably not optimal. The model involved numerous variables which would necessarily be set equal to constants chosen by the examiners prior to implementation of the tailored testing procedure.

An interactive computerized approach to tailored testing seems appropriate and, in many cases, necessary. The impracticality of a computer terminal for each examinee, as Wood (1973) has indicated, has been realized. But the cost of computerized testing may not be high in relation to the costs incurred in administering group tests of any extensive length or of extent in number, especially if the costs of scoring, analysis, and reporting are included. Also, with the increased emphasis on individualized instruction, in which frequent measurement of student performance is necessary, it seems that an interactive computerized testing procedure is desirable. Preferred features of such a system might include (1) reduction of testing time with minimal loss of information, (2) maintenance of records of student performance, and (3) indications of appropriate actions to be taken for the benefit of the student.

Excluding the factor of cost, the tailored testing model proposed by Kalisch (1974) may be practical for instructional testing only if the following provisions are possible:

1. The number of Ss from whom responses to the items would be necessary prior to implementation of the tailored testing

procedure is not too large.

2. The decisions to be made by the examiners in terms of values that must be specified for variables within the testing model are neither too difficult nor too numerous.

As a means of meeting the second criterion, two adaptations of Kalisch's (1974) model are presented in this paper. Each of these adaptations eliminates the need for some of the decisions that were necessary prior to implementation of the earlier testing model. The study sought to determine if the new models are practical, in terms of the first criterion specified above, and also valid, in terms of accurate prediction of the correctness of examinees' item responses.

The study employed computer simulations, using item response data previously collected in 1973, through the Florida Statewide Assessment Program.¹ The data used were part of a statewide objective-referenced testing program in the area of mathematics for ninth grade students.

¹Florida Statewide Assessment Program, 1973, State Department of Education, Tallahassee, Florida.

CHAPTER II

REVIEW OF RELATED LITERATURE

Research regarding tailored testing has primarily been conducted in the area of aptitude and achievement tests, with regard to the measurement of underlying traits and not with regard to instructional testing (Cleary, Linn, & Rock, 1968a; Hubbard, 1966; Linn, Rock, & Cleary, 1969; Lord, 1971; Waters & Bayroff, 1971). Instructional applications of tailored testing have been pursued by the following individuals:

1. Ferguson (1970), who investigated an instructional tailored testing procedure involving a hierarchical arrangement of objectives.
2. Green (1970), who suggested the use of instructional tailored testing for dichotomous decision making procedures.
3. Linn, Rock, and Cleary (1972), who validated Green's theory concerning the efficiency of making dichotomous decisions with tailored testing.

Purposes of Tailored Testing

Lord (1970) has indicated that the purpose of a tailored test is to obtain optimal measurement of an individual's ability, whereas Cleary, Linn, and Rock (1968a) have indicated a more

general purpose--to present to an examinee items which are appropriate for his level of performance. Lord (1971) demonstrated that the tailored testing procedure he used was primarily of value in testing examinees for whom the normal tests would be too hard or too easy.

Green's (1970) purpose in making a dichotomous decision in an instructional situation with the use of a tailored test was to reduce the number of items of a testing instrument. Green showed that, although a flattening effect occurs in the test characteristic curves as the number of test items decreases, a decision may be made with a 20- or 30-item test rather than a 60-item test.

Using a sequential decision procedure, the number of items necessary for a decision varies as the level of the ability of the examinee, with the greater number of items necessary for those at the mean ability level. Linn, Rock, and Cleary (1972) validated Green's theory concerning the efficiency in making dichotomous decisions with tailored testing. Conventional tests required approximately twice as many items as sequential tests to produce the same accuracy in decision making.

Ferguson's (1970) purpose for using tailored testing was to obtain sufficient information on the basis of an examinee's responses in order to decide whether the examinee had or had not mastered instructional objectives. If insufficient information was available to make a decision concerning the mastery of an objective by an examinee, another item corresponding to the objective was presented to the examinee.

Assumptions

The assumptions underlying all tailored testing models are, of course, not the same. Tailored testing procedures employing the normal ogive model (Lord & Novick, 1968) and Birnbaum's three-parameter logistic function (Birnbaum, 1968; Lord, 1968) assume a unidimensional underlying trait. Lord (1968) has agreed that much research will be necessary to determine the appropriateness of the assumption, but "it may provide a tolerably good approximation [p. 990]" for certain tests.

For tailored testing purposes, test items have generally been scored as either "correct" or "wrong." Wood (1973) has suggested that greater information be obtained from "wrong" responses, using methods such as those by Bock (1972) and Samejima (1969). Green (1970) has argued that present theory uses a right-wrong scoring that fails to consider the elaborate mental processes which are required in responding to an item. The collection of such data is possible in a computer-based situation. Green has suggested that the sequence of selection of alternative answers to an item may prove to be more useful than dichotomous scoring of items.

Hansen (1969) investigated the use of response confidence ratings in predicting total scores on a computer-presented tailored test. Hansen also suggested that response latencies may be as meaningful as confidence ratings. Anxiety status may also be considered as a variable in determining the sequence of the items to be presented.

The effects of guessing by the examinee have been considered by Birnbaum (1968) by employing the three-parameter logistic model. Use of methods that determine the examinee's sequence of selection of alternative answers, anxiety, and confidence of an examinee's response, as discussed previously, might also be applicable to the determination of guesses.

Ferguson's (1970) tailored testing model for determination of mastery or non-mastery of objectives assumes a valid hierarchy of the objectives. The model also assumes that the examinee's mastery of objectives may be determined by the examinee's performance at his final level in the hierarchy. This assumption corresponds to the assumption in the two- and three-parameter logistic models that the examinee's ability level is characterized in terms of the difficulty levels of the items he can answer (Lord, 1970). In an instructional situation, it is generally neither feasible nor desirable to pinpoint an individual's performance on a continuum. The approach is probably not feasible unless the tasks specified by the objectives are linearly hierarchically dependent and the ranges of difficulties of the items within objectives do not intersect. Even if a valid hierarchy exists for a set of cognitive skills, the hierarchy is not necessarily linear.

Branching Strategies

A tailored test consists of two parts: (1) a scoring section, in which the examinee's response is scored and related measures are obtained, and (2) a branching, or routing section. In the models

investigated by Lord (1970, 1971) branching was directed toward an item of greater or less difficulty, depending upon the examinee's respective correctness or incorrectness of his response. Cleary, Linn, & Rock (1968b) presented items in descending order of the magnitude of the point-biserial correlations with the total test score.

Ferguson (1970) found that his tailored testing procedure was most efficient if items whose objective was at the midpoint of the hierarchy were presented first. Branching was then directed to the midpoint between the initial objective tested and the most difficult or easiest untested objective, depending upon whether the student exhibited or failed to exhibit mastery of the objective, respectively. Branching to a more difficult objective implied mastery of the lower order objectives, without actual testing on the objectives.

Branching strategies often incorporate step sizes. In Lord's models in which item difficulties were used, two types of stepping methods have been identified: (1) fixed step size and (2) shrinking step size (the Robbins-Monroe procedure). In terms of item difficulties, a fixed step size approach routes the examinee to an item with a difficulty value of $d \pm i$, where d equals the difficulty of the previous item and i is the fixed step value. A shrinking step size requires reduction in the size of i each time an item has been answered by the examinee. It would appear that shrinking step sizes would permit rapid matching of item difficulty to an

examinee's ability in the early stages of a test and close matching afterwards, when the examinee's level might accurately be inferred on the basis of his previous responses (Lord, 1971).

Although a shrinking step size suggests greater efficiency in tailored testing, Lord (1971) has shown that this is not true under all conditions. In terms of a scoring system which requires determination of the difficulty of item $n+1$, and in terms of a shrinking step size, approximating a harmonic sequence, the fixed step size and the Robbins-Monroe procedures provided approximately equal measurement. Since the latter method requires the availability of 2^n items for an n -item test, the fixed step method appears to be more efficient.

Scoring

In order to use a tailored testing operation, a scoring system which is dependent upon the type of tailored test and the purpose of the test, must be used. Lord (1970) discussed three possible scoring schemes, appropriate to the latent trait models. The scoring methods were: (1) number right score, (2) final difficulty score, and (3) average difficulty score. The number right scoring scheme may be appropriate if all examinees are to receive the same number of items. In the tailored testing procedures used by Ferguson (1970) and Linn, Rock, and Cleary (1972), an examinee was given additional items until it was possible to classify him as a member of one of two groups. Even if the use of item difficulty is appropriate for scoring purposes, the use of

average difficulty scores and final difficulty scores may be differentially effective (Lord, 1971) according to the number of items presented, test item characteristics, and abilities or achievement levels of the examinees.

Sizes of the Measurement Sections

Prior to a branching decision in a tailored testing procedure, a response to at least one item must be obtained from the examinee. If a branching decision is to be made after the examinee's first response, the effectiveness of the test may be highly dependent upon the difficulty level of the first item and the probability of the examinee's randomly selecting an answer (Lord, 1971). A branching decision based upon a random guess by an examinee may direct the examinee to items inappropriate to his true ability or performance level. Of course, prior to an initial or subsequent branching decision, responses to more than one item may be obtained. Waters and Bayroff (1971) compared the results of tailored testing procedures using one item per stage and two items per stage branching strategies. Their results showed that the one item per stage strategy produced higher correlation coefficients between tailored test scores and ability levels of the examinees than did a two item per stage strategy. Waters and Bayroff used the same two-parameter latent trait model employed by Lord. A tailored testing procedure employing another model or using another criterion for comparing the effectiveness of tailored testing and conventional testing may indicate that multi-item per stage strategies are more

effective than a one item per stage strategy. Also, although a one item per stage strategy might generally be a superior branching strategy, a multi-item measurement section prior to the first branching decision, followed by subsequent one item per stage or a variable number of items per stage procedure might be effective.

Kalisch's Tailored Testing Model

As a means of applying tailored testing to instructional measurement, Kalisch (1974) proposed a model which may be applicable to proper placement of individuals in an instructional sequence and serve to evaluate an individual's performance after instruction. The proposed model may also reduce testing time in that it provides a procedure for presenting a minimum number of items of a test and predicting the correctness of the responses to the remaining items. Kalisch (1974) has contended that the model is not dependent upon an assumed or validated hierarchy of skills and that the necessary number of responses to items by Ss, prior to implementation of the model, would be sufficiently small to provide accurate predictions and a practical means of successfully employing computer-based tailored testing. The model employs item conditional difficulties, beta distributions, and an expected loss function.

Kalisch's (1974) model makes the following assumptions:

1. The probability of the correctness of an examinee's responses to items (item conditional difficulties) may be estimated by analyzing the response patterns of individuals who have

previously answered the items and who have displayed the same item response pattern as the examinee.

2. The distribution of sample conditional difficulties of a specified item is approximated by a corresponding beta distribution, whose mean equals the difficulty of an item and whose variance is approximately equal to the sampling variance of the item conditional difficulty.

The model requires that before any predictions concerning the correctness of an examinee's response to an item are made, at least one item must be presented to the examinee. The item or items presented prior to any attempt at prediction are those with difficulty values closest to 0.5. Subsequently, the branching strategy employed requires the presentation of the item whose conditional difficulty for the examinee is closest to 0.5. Each decision in the branching strategy is dependent on all the previous responses given by the individual. Since the conditional difficulties of the unrepresented items for the examinee may and probably do change as he answers additional items, the branching strategy is not easily categorized as being one that uses either a fixed or a shrinking step size, or even a combination of the two.

After the presentation of the initial items (those presented prior to any attempt to predict the correctness of the remaining item responses) and after the subsequent presentation of any other items, one of three decisions is made for each item for which no previous decision to predict the examinee's response has been made.

The three decisions are:

1. Predict that the examinee would answer the item correctly.
2. Predict that the examinee would answer the item incorrectly.
3. Make no prediction concerning the examinee's response to the item.

For each unpresented item for which no decision to predict response correctness has been made, probabilities are computed for three information categories: $\rho < V_1$, $V_1 \leq \rho \leq V_2$, and $\rho > V_2$, where V_1 and V_2 are the critical conditional difficulty values, supplied by the examiner. The probabilities are determined, using the related beta distribution, based upon the total number of responses to the item. The probability for each information category is used as a weighting of the loss value assigned to each information category-by-decision cell. The sum of the weighted losses within a decision category provides a measure of the expected loss for the decision. Decision theory (Hays & Winkler, 1970) indicates that the decision with the minimum expected loss is the best decision.

In the model proposed by Kalisch (1974), for each item, the item response is obtained from the examinee or the correctness of his response is predicted. As additional items are presented to S, the number of observations in the data base upon which conditional probabilities are computed is reduced. If at least two observations are not available, the remaining items for which responses have neither been obtained nor predicted, are presented to the S for his responses. The stopping rule for the model is that, for

each item, a response has been obtained from the examinee or a decision has been made to predict the correctness of his response, had the examinee supplied one.

Scoring is in terms of item scores. Although Kalisch has indicated that the tailored testing model might be applicable in determining mastery of instructional objectives, a specific procedure for translating item scores to decisions concerning mastery of corresponding objectives has not yet been advanced.

Although the validation study reported by Kalisch (1974) required, on the average, Ss to respond to eight items and produced correct predictions for 10 of the remaining 12 items of the test, the results were probably not optimal. Investigation of the validity of the model has been minimal and a study of optimizing the model is non-existent. Further refinement of the model and investigations of the variables within the model seem worthwhile, since the model may provide a relatively simple and usable tailored testing procedure.

CHAPTER III

THE TAILORED TESTING MODELS

The model proposed by Kalisch (1974) involved numerous variables which would necessarily be set equal to constants chosen by the examiners prior to implementation of the tailored testing procedure. Generally, the fewer the decisions to be made by the examiners the greater the ease in implementation. Hence, a revised version of Kalisch's earlier model is suggested. The revised model eliminates the need for some of the information previously required. The revised model will henceforth be referred to as Model 1.

In both the earlier model and Model 1, the decisions concerning prediction of the correctness of the examinee's item responses have not been reconsidered, even though additional information is obtained whenever the examinee answers an additional item. An adaptation of Model 1, hereafter referred to as Model 2, is identical to Model 1 with the exception that as additional information is obtained, previous decisions are reconsidered.

As a preface to the discussion of the tailored testing models, an explanation of some of the relevant terminology follows.

Conditional Difficulty of an Item

The proposed model uses a branching strategy and a decision making procedure that require determination of the probability that

an examinee will answer a specified item correctly. The item difficulty provides such a probability value, but fails to make use of the information concerning the correctness of the examinee's responses to the previously presented items. Use of such information, although impractical, if not impossible, in paper and pencil situations, is possible in computer-based situations.

The conditional difficulty of an item for an individual is the probability of the individual's answering the k -th item correctly, given information concerning the correctness of the individual's responses to the preceding $k-1$ items and given prior subject-item response data. Hence, a conditional difficulty is a conditional probability. Let C_i represent the event that item i was answered correctly; W_i represent the event that item i was answered incorrectly; and $P(B|A)$ represent the event B given event A . Table 1 contains a subject by item matrix of ones and zeros representing correct and incorrect responses, respectively. If an individual has answered the first item correctly, the conditional difficulty of the remaining two items is determined as follows:

$$P(C_2|C_1) = \frac{P(C_2 \cap C_1)}{P(C_1)} = \frac{5}{8} \quad [1]$$

$$P(C_3|C_1) = \frac{P(C_3 \cap C_1)}{P(C_1)} = \frac{4}{8} \quad [2]$$

If the individual correctly answered the first item but incorrectly answered the second item, the conditional difficulty of the third item is

$$P(C_3|C_1, W_2) = \frac{P(C_3 n C_1 n W_2)}{P(C_1 n W_2)} = \frac{2}{3}. \quad [3]$$

TABLE 1

Subject by Item Response Matrix
for Ten Subjects and Three Items

Subjects	Items		
	1	2	3
1	1	1	0
2	1	1	1
3	1	1	0
4	1	1	1
5	1	1	0
6	1	0	1
7	0	1	0
8	1	0	1
9	0	0	0
10	1	0	0

Item Conditional Difficulty and the Beta Distribution

The tailored testing models being presented use population item conditional difficulties. Since the values obtained for the conditional difficulty of an item may vary according to the samples of examinees selected, the model considers the sampling variance of the conditional difficulty of an item. For any item with a sample conditional difficulty value between zero and one, there exists a beta distribution which has the following properties:

1. The expected value of the distribution equals the sample item conditional difficulty.

2. Its variance is nearly equal to the sampling variance of the conditional difficulty of the item.

3. Its domain is the closed interval from zero to one. It has therefore been assumed that there exists a beta distribution which approximates the distribution of item conditional difficulty values obtained from infinitely many samples of examinees.

Consider a dichotomously scored item answered by N individuals, r of whom answered the item correctly. An unbiased estimate of the population parameter of item difficulty is $p=r/N$. For an item (Hays & Winkler, 1970), the sampling variance s^2 is given by the equation

$$s^2 = \frac{Nr-r^2}{N^2} = p(1-p). \quad [4]$$

An unbiased estimate of the population variance σ^2 is given by the equation

$$\hat{\sigma}^2 = \frac{N}{N-1} s^2 = \frac{Nr-r}{N(N-1)} = \frac{Np(1-p)}{N-1}. \quad [5]$$

The variance of the sampling distribution of item conditional difficulties σ_M^2 is given by the equation

$$\hat{\sigma}_M^2 = \frac{\hat{\sigma}^2}{N} = \frac{r(N-r)}{N^2(N-1)} = \frac{p(1-p)}{N-1}. \quad [6]$$

The beta distribution (Hays & Winkler, 1970), specified by the equation

$$\beta(x) = \begin{cases} \frac{(N-1)!}{(r-1)!(N-r-1)!} x^{r-1} (1-x)^{N-r-1} & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x < 0 \text{ or } x > 1 \end{cases} \quad [7]$$

is defined in terms of the random variable x , the probability of a "success" on any single Bernoulli trial. The mean of the distribution is $p = r/N$, and the variance of the beta distribution σ_B^2 is given by the equation

$$\sigma_B^2 = \frac{r(N-r)}{N^2(N+1)} = \frac{p(1-p)}{N+1}. \quad [8]$$

Examples of beta distributions are shown in Figure 1. A beta distribution approximating the assumed distribution exists only if $N \geq 2$.

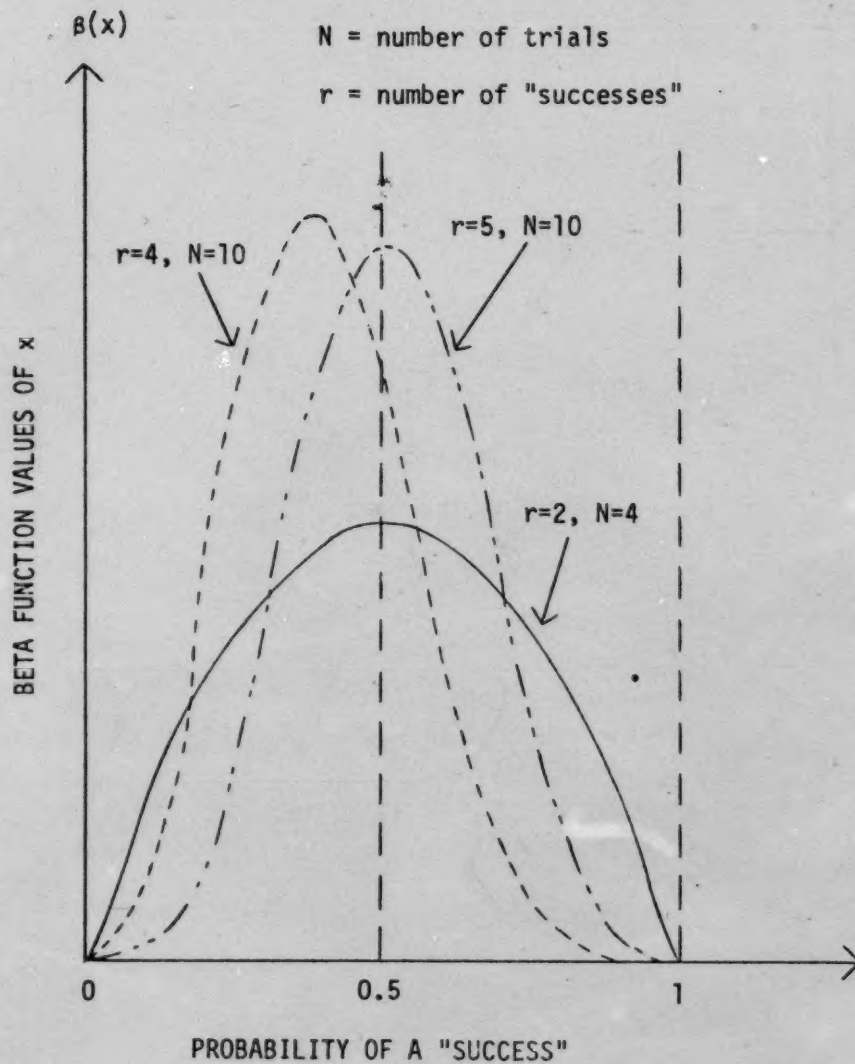


Fig. 1. Three beta distributions.

The models investigated use loss functions weighted by the probabilities of the population conditional difficulties of items being at each value in the closed interval $[0,1]$. Such probabilities may be determined by using a beta distribution. Since the area of the region between the horizontal axis and a beta distribution curve is one unit, the probability that the conditional difficulty of an item is in the interval $[a,b]$ is calculated from the formula

$$\begin{aligned}
 P(a \leq x \leq b) &= \int_a^b f(x) dx \\
 &= \int_a^b \frac{(N-1)!}{(r-1)!(N-r-1)!} x^{r-1} (1-x)^{N-r-1} dx. \quad [9]
 \end{aligned}$$

Loss Functions

The tailored testing procedures being investigated require the prediction of the correctness of an examinee's responses to items, based upon his responses to previously presented items. If the population conditional difficulty of an item for an individual were equal to one, the decision maker should be willing to assume that the examinee would answer the item correctly if it were presented. A conditional difficulty equal to zero would warrant a decision to assume the examinee's response to be incorrect. For population conditional difficulties between zero and one, the decision to predict the correctness of the examinee's response becomes a more difficult task.

Table 2 depicts a loss matrix applicable to the tailored testing models used. The dimension, information categories (Cronbach & Gleser, 1965), specifies two cases involving the population conditional difficulty ρ of an item: $\rho=1$ and $\rho=0$. The second dimension, the action, is separated into three levels: (1) predict the examinee's response to be correct, (2) predict the examinee's response to be wrong, and (3) obtain the examinee's response to the item.

TABLE 2

Matrix of Loss Values for Items with
Conditional Difficulty Values of 1 or 0

Action	Information Categories	
	$\rho = 1$	$\rho = 0$
Predict examinee's response to be correct	L_1	L_2
Predict examinee's response to be wrong	L_3	L_4
Obtain examinee's response to the item	L_5	L_6

The six values (represented by L_i 's, $i=1,6$) in the matrix are the expected losses for a specified action, given a particular information category (Hays & Winkler, 1970).

A matrix such as depicted in Table 2 may be used in a decision making situation under one of two conditions--certainty or uncertainty (Hays & Winkler, 1970). Under a condition of certainty, the decision maker knows the truth concerning the information categories, whereas under a condition of uncertainty, the actual truth is unknown. For example, in Table 2, a condition of certainty would exist if the decision maker actually knew whether the conditional difficulty were either zero or one. In a condition of uncertainty, the decision maker might have information which would imply the value of p , but the actual value of p would not be known to him. In either case an appropriate action could be determined, although quite naturally, more easily under a condition of certainty than uncertainty.

The author envisions two types of loss functions that may be useful in education. The first loss function, either under a condition of certainty or uncertainty, may be used for decision making purposes, such as predicting the correctness of item responses. The second type of loss function is envisioned as being used for the purpose of evaluation. Consider the type of tailored testing procedure under discussion. An evaluation would probably indicate that the model is effective if (1) relatively few items are presented to each student and (2) numerous, possibly all, predictions of the correctness of student's responses to items

are accurate. The presentation of too few items to an examinee might provide insufficient information for many correct predictions and produce numerous incorrect predictions. Of course, no incorrect predictions will occur if all the items are presented to the examinee. A loss matrix as shown in Table 3 may be employed to evaluate the tailored testing procedure or compare different versions of the procedure.

TABLE 3
Matrix of Loss Values
for Determining Actual Loss

Action	Item Response, if Obtained, Would Be	
	Correct	Wrong
Predict examinee's response to be correct	L_1	L_2
Predict examinee's response to be wrong	L_3	L_4
Obtain examinee's response to the item	L_5	L_6

The values of the loss matrix may incorporate such factors as the following:

1. Instructional costs.
2. Amounts of time for instruction.
3. Effects on the students' morale.

In a computer-based testing situation, a greater loss may be attributed to the increased time and cost needed in presenting an examinee an item and processing his response than in accurately predicting the correctness of his response. Hence, it would be anticipated that the losses indicated by L_1 and L_4 in Table 3 would be less than L_5 and L_6 . In terms of the instructional costs and the inappropriate use of the student's time that might result from incorrect placement, resulting from incorrect prediction and from the ensuing corrective actions, it is anticipated that L_2 and L_3 would be greater than L_1 and L_4 . A specific ordering of the loss values, applicable to all situations, is not possible, since the specification of the values is dependent upon the purpose of the test, the factors considered, and the decision makers. Although the author believes that in most instructional situations greater losses would be associated with incorrect predictions (L_2 and L_3) than with obtaining item responses (L_5 and L_6), a reverse ordering is possible.

The loss values, indicated by L_i 's ($i=1,6$) in Table 3, are specified according to the evaluator's criteria. A tailored test of the type being considered might be administered to a sample

of examinees and after all decisions had been made (predictions of the correctness of item responses) for a student, his responses to those items would then be obtained. Hence, given the student's actual response and the action taken for each item during the tailored testing sequence, the actual loss can be determined. The total actual loss for an examinee might be defined as the sum of the item losses. A loss matrix such as in Table 3 would therefore provide a quantified scheme for evaluation.

A Linear Loss Assumption

The tailored testing models involved in this study require the use of loss functions under conditions of uncertainty in decision making situations. An assumption in the models is that if the population conditional difficulty of an item is one, the examinee will answer the item correctly. Similarly, if $p=0$, it is assumed that the examinee will answer the item incorrectly. For the values of p between zero and one, the models assume linear loss functions, one within each of the two actions involving prediction. Figure 2 depicts two linear loss functions, applicable to the tailored testing models.

A Non-Linear Loss Assumption

Although a linear loss assumption may be appropriate for the decisions involving prediction of a response, the same assumption does not seem appropriate for determining the loss associated with obtaining the S's response. If the conditional difficulty of an item is either one or zero, the expected loss for obtaining a response

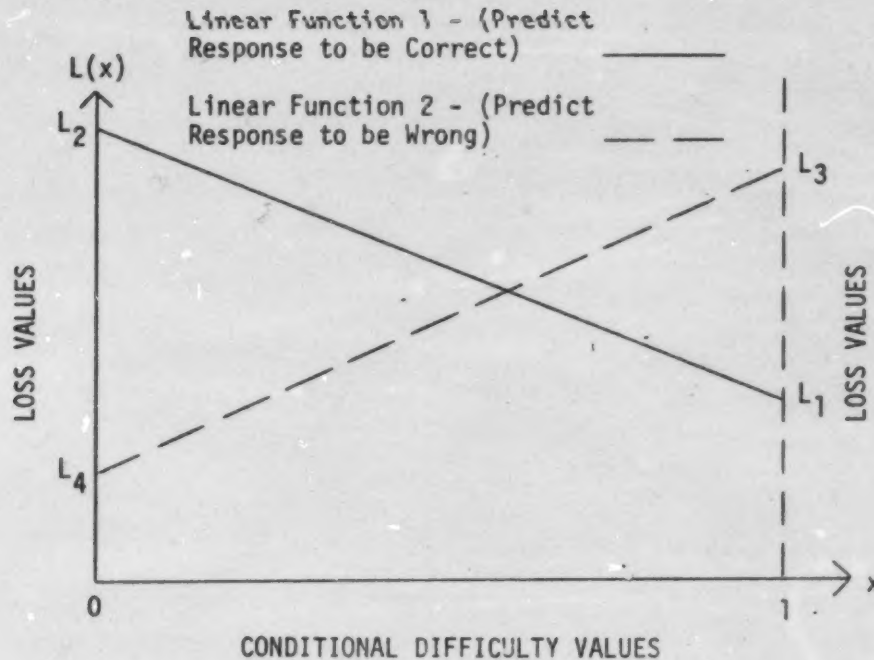


Fig. 2. Two Linear Loss Functions.

would be expected to be higher than the losses associated with a prediction of the response. For items with conditional difficulties closer to the value of 0.5, it would be expected that the loss for obtaining a response would be less. A loss function appropriate for the action, obtain the student's response, appears to be a function that is decreasing in the interval $[0, 0.5]$ and increasing in the interval $[0.5, 1]$. A function meeting these conditions is the quadratic function with an equation of the form $L(x) = k(x - 0.5)^2$, where $k > 0$. Figure 3 shows the non-linear loss function described. Its use in conjunction with beta distributions, in order to determine expected loss, is to be discussed later.

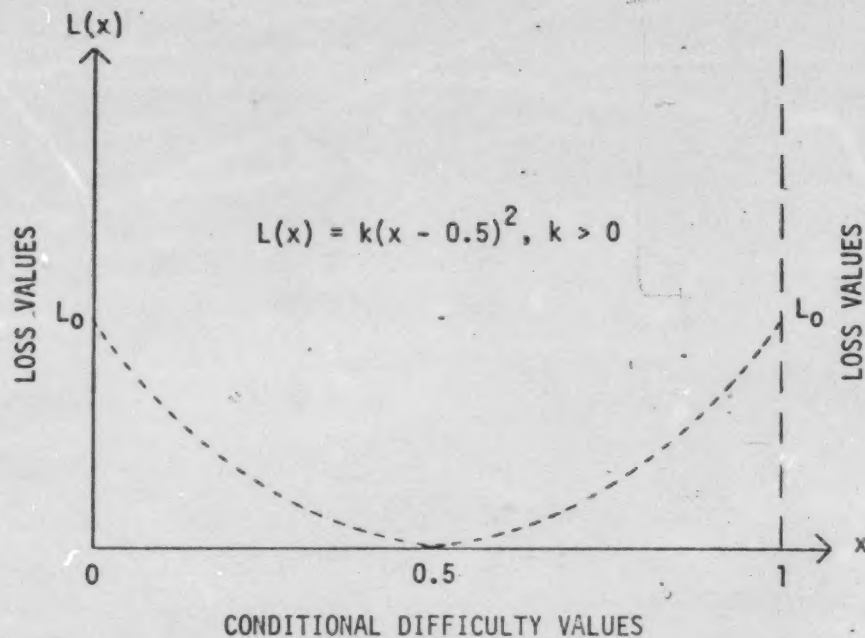


Fig. 3. A non-linear loss function.

Using an Evaluation Loss Matrix for Decision Making

Two types of loss matrices have been identified. One type of loss matrix may be used for evaluation of a process and the other type may be used for decision making within the process. The question that naturally arises is whether the same loss matrix may be used for both the evaluation and decision making phases of a process. In relation to the tailored testing models used in the study, a preliminary investigation of this dual use of a loss matrix was conducted. Three different loss matrices, designed as evaluation loss matrices, were used in conjunction with beta distributions, to compute expected losses within each of the three

actions. The results indicated a high proportion of incorrect predictions. For values of the conditional difficulty practically at 0.5, the procedure yielded predictions. From an intuitive viewpoint, it was judged that the expected loss for obtaining the examinee's response was very often too high, in relation to the expected loss for the other two actions.

A corrective action that one might consider is to increase the loss values associated with the actions involving prediction of the examinee's response. In fact, it is possible to increase the loss values for the actions involving prediction to guarantee that no prediction is made unless the estimated population conditional difficulty of an item is greater or less than certain prescribed values. In the preliminary investigation, loss values were selected so that the decisions to predict within the tailored testing procedure produced substantially fewer incorrect predictions than previously. But when the same loss matrix was used for evaluation, the total loss for a student increased exorbitantly whenever an incorrect prediction had been made. So great was the total loss for any case in which the incorrect predictions had been made that the total loss exceeded what the loss would have been if all the items had been presented to the examinee.

The conclusion drawn from the preliminary study was that the loss functions used within the tailored testing model neither need be the same nor probably should be the same as the loss functions used in the evaluation.

Determining Expected Loss

As in Kalisch's (1974) model, the sample conditional difficulty of an item, p , is used as an estimate of the population parameter ρ . For each determined conditional difficulty between zero and one, there is a beta distribution whose expected value equals the estimated population conditional difficulty and whose variance closely approximates the sampling variance of the conditional difficulty. Within each action the loss value corresponding to each point in the interval $[0,1]$ is weighted by the corresponding probability, determined from the beta distribution, and the weighted loss values are summed, to obtain the expected loss for the action. For all $N \geq 2$ and all r such that $0 < r < N$ there exists a beta distribution, β , as specified in Formula 7. Also, if for an action and for each value of x in the open interval $(0,1)$ there exists a loss value $L(x)$, then the expected loss $E(A_i)$ for the action A_i , is given by the formula

$$E(A_i) = \int_0^1 L(x)\beta(x)dx. \quad [10]$$

Let L be a linear loss function defined by the equation

$$L(x) = L_i + (L_j - L_i)x, \quad [11]$$

where L_i and L_j are the loss values for an action for the conditions $\rho=0$ and $\rho=1$, respectively. For example, in Table 2, for the action of predicting an examinee's response to be wrong, L_i equals L_4 and L_j equals L_3 . Substituting Formulas 7 and 11 for $\beta(x)$ and $L(x)$, respectively, in Formula 10, a resulting formula is

$$E(A_i) = \int_0^1 [L_i + (L_j - L_i)x] \frac{(N-1)!}{(r-1)!(N-r-1)!} x^{r-1} (1-x)^{N-r-1} dx. \quad [12]$$

By rearranging terms and using integration by parts, Formula 12 may be reduced to the following form:

$$E(A_i) = \frac{1}{N} [(N-r)L_i + rL_j]. \quad [13]$$

The reduction of Formula 12 to Formula 13 is shown in Appendix A. Formula 13 is applicable for determining the expected loss for the actions involving prediction.

For the decision, to obtain the examinee's response, the loss function employed is the quadratic relation of the form

$$L(x) = k(x - 0.5)^2, \text{ where } k > 0. \quad [14]$$

Let L_i and L_j be the loss values for the action for the conditions $p=0$ and $p=1$, respectively. Assume also that $L_i = L_j$. Hence, Formula 14 becomes

$$L(x) = 4L_i(x - 0.5)^2. \quad [15]$$

Substituting Formulas 7 and 15 for $\beta(x)$ and $L(x)$, respectively, in Formula 10, a resulting formula is

$$E(A_i) = \int_0^1 [4L_i(x-0.5)^2] \left[\frac{(N-1)!}{(r-1)!(N-r-1)!} \right] x^{r-1} (1-x)^{N-r-1} dx. \quad [16]$$

By rearranging terms and using integration by parts, Formula 16 may

be reduced to

$$E(A_i) = L_i \left[\frac{4r(r-N)}{N(N+1)} + 1 \right] \quad [17]$$

The reduction of Formula 16 to Formula 17 is shown in Appendix B.

Selecting Loss Values for the Decision Loss Matrix

As previously discussed, the criteria to be used in determining the loss values for a matrix intended for evaluative purposes may include instructional costs, amounts of time for instruction, and the effects on students' morale. Also, as previously mentioned, a loss matrix to be used for evaluation of a process may not be adequate as a decision making device within that process. Since the tailored testing models employ loss matrices, and loss matrices designed for evaluating the models were shown to be inadequate for decision making within the models, an analytical means of determining the needed loss values was used.

Kalisch's (1974) model required the decision maker to preset the critical conditional difficulty values V_1 and V_2 , which specified three possible conditions: $\rho < V_1$, $V_1 \leq \rho \leq V_2$, and $\rho > V_2$. These three conditions provided the decision maker with three discrete information categories for the decision loss matrix. The linear loss assumptions for the actions involving prediction and the quadratic loss assumption for the action, to obtain a response, have eliminated the need for the decision maker to supply both the critical conditional difficulty values and the loss values for the

decision loss matrix. The loss values may be expressed as a function of the critical conditional difficulty values.

Consider the expected loss for predicting the correctness of a response as determined from Formula 13. An equivalent form of Formula 13 is

$$EL(A_i) = \frac{N-r}{N} L_i + \frac{r}{N} L_j. \quad [18]$$

Since the conditional difficulty of an item is $p = r/N$, Formula 18 may be written as

$$EL(A_i) = (1-p)L_i + pL_j. \quad [19]$$

Since L_i and L_j are constants, for any given value of p , the expected loss will be the same regardless of the actual values of r and N .

Hence, if $p = 0.3$, the expected loss for predicting a response is the same for sample sizes of 10, 100, 1000, or any other value of $N \geq 2$.

Such constancy in expected loss values for given values of p is not the case for the expected loss for obtaining a response. The formula for expected loss for obtaining a response, as specified in Formula 17, may be expressed in an alternate form,

$$EL(A_i) = L_i \left[4 \left(\frac{r}{N} \right) \left(\frac{r-N}{N+1} \right) + 1 \right]. \quad [20]$$

Substituting p for r/N in Formula 20, the resulting formula is

$$EL(A_i) = L_i \left[4 \frac{p(p-1)}{1 + \frac{1}{N}} + 1 \right]. \quad [21]$$

The expected loss for obtaining a response is dependent not only upon the value of p but also upon the sample size N . From Formula 21, it is noted that

$$\lim_{N \rightarrow \infty} EL(A_i) = L_i \left[4(p^2 - p) + 1 \right]. \quad [22]$$

In both Formulas 21 and 22, the loss constant L_i is multiplied by a factor. Table 4 lists values of the factors for specific values of the conditional difficulty and for various values of N . From Table 4 and Formulas 21 and 22, it is apparent that for conditional difficulty values other than zero and one, the expected loss for obtaining a response decreases for a given conditional difficulty as N increases. As N increases and the conditional difficulty approaches 0.5, the value of the factor approaches zero. Intuitively, this result seems quite appropriate for predictive purposes since for a given conditional difficulty, the expected loss decreases as the size of the data base increases. Loss for this decision is minimized when the data base consists of responses from the entire population and the conditional difficulty equals 0.5.

In order to determine the loss values for the decision loss matrix, the following additional assumptions have been made:

1. The loss value for predicting a response to be correct when $p = 1$ equals zero.
2. The loss value for predicting a response to be wrong when $p = 0$ equals zero.
3. The loss value for predicting a response to be wrong when

TABLE 4
Values of Factors for Formulas 21 and 22

Conditional Difficulty	N = 2	N = 10	N = 2000	N → ∞
0 or 1	1.000	1.000	1.000	1.000
.1 or .9	0.763	0.673	0.642	0.640
.2 or .8	0.573	0.418	0.363	0.360
.3 or .7	0.443	0.236	0.164	0.160
.4 or .6	0.363	0.127	0.045	0.040
.5	0.333	0.091	0.005	0.000

$p = 1$ equals the loss value for predicting a response to be correct when $p = 0$.

Figure 4 depicts the corresponding loss functions with respect to these additional assumptions. Using the notation L_p for the critical loss value for prediction, the notation L_o for the critical loss value for obtaining a response, and the three additional assumptions, Formulas 19 and 21 for expected loss may be written in the following forms:

1. For predicting a response to be correct,

$$EL(PC) = (1-p)L_p \quad [23]$$

2. For predicting a response to be wrong,

$$EL(PW) = pL_p \quad [24]$$

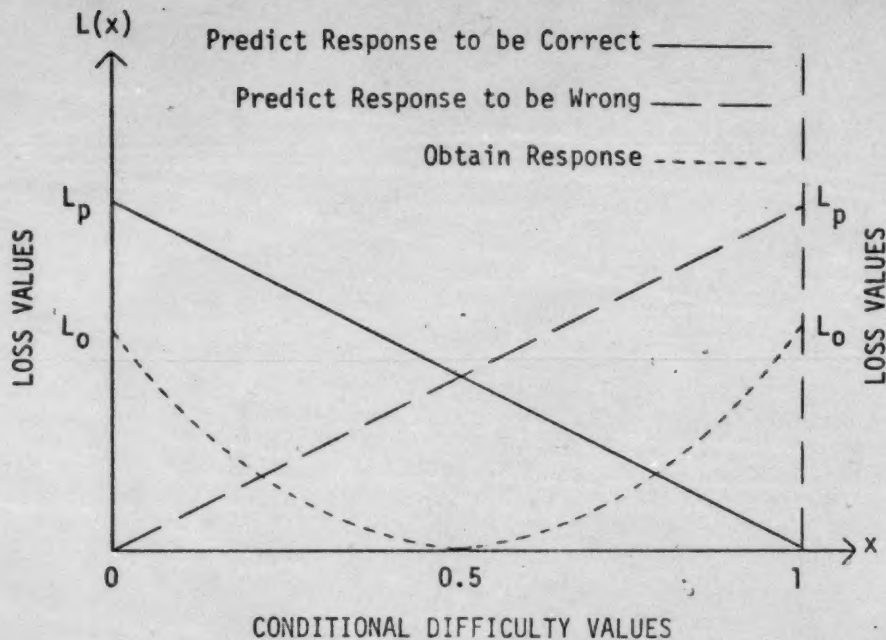


Fig. 4. The three loss functions used in the study.

3. For obtaining a response,

$$EL(OR) = L_o \left[\frac{4p(p-1)}{1 + \frac{1}{N}} + 1 \right]. \quad [25]$$

One must now determine the values of L_p and L_o . Since the data base will not have responses from the entire population, it is necessary to specify the values of p such that $EL(PC) < EL(OR)$ and $EL(PW) < EL(OR)$. In other words, one must determine critical conditional difficulties such that the expected loss for predicting the correctness or incorrectness of a response is less than the expected loss for obtaining a response. The theory underlying the tailored testing models does not provide a means of determining the

critical conditional difficulty values. Hence, the decision maker must supply these values. Under the present assumptions, the decision maker need only select one critical conditional difficulty value, V . The other critical value equals $1 - V$.

As previously mentioned, the expected loss for predicting is not dependent upon the sample size N , whereas the expected loss for obtaining a response is dependent upon N . Since the expected loss for obtaining a response to an item with a given conditional difficulty increases as the sample size decreases, the decision maker must specify the value of V and the sample size of the data base to which the critical conditional difficulty value corresponds.

Suppose that the decision maker decides to select $V = 0.8$ in relation to a data base of size $N = 100$. For values of $p > V$, the decision maker desires the expected loss for predicting a response to be correct to be less than the expected loss for obtaining a response. Hence, from Formulas 23 and 25, $EL(PC) < EL(OR)$ implies

$$(1 - p) L_p < L_o \left[\frac{4p(p-1) + 1 + \frac{1}{N}}{1 + \frac{1}{N}} \right]. \quad [26]$$

Substituting 0.8 for p and 100 for N in Formula 26, the following relation is obtained

$$L_p < L_o (1.832). \quad [27]$$

Formula 27 specifies the conditions under which the loss for

predicting a response as correct is less than the loss for obtaining the response. The values for L_p and L_o are relative. For convenience, let $L_o = 100$. Then $L_p = 183$ (to the nearest whole number). Hence, by specifying the critical conditional difficulty value of V and a base size N , values for L_p and L_o have been determined.

Table 5 shows the expected loss values for obtaining a response for various data base sizes, under the conditions $L_p = 183$ and $L_o = 100$. The expected loss for predicting a response to be correct for an item conditional difficulty value of 0.8 is 36.6, regardless of the size of the data base. However, as Table 5 indicates, the expected loss for obtaining a response is dependent upon the size of the data base. If the decision is not to predict unless the conditional difficulty is greater than 0.8, regardless of the size of the data base, the value of L_p and L_o would be computed on the basis of $N = 2$, the smallest possible size of a data base.

Table 6 shows adjusted critical conditional difficulty values for a selected critical conditional difficulty of 0.8, corresponding to a specified data base size, denoted by N_D . For example, if the decision maker selected the critical conditional difficulty of 0.8 for a data base size of $N_D = 10$, the adjusted critical conditional difficulty for a data base size two would be 0.757.

Model 1

The first tailored testing model that was investigated is specified by the following procedure:

1. A data base consisting of a subject by item matrix of

TABLE 5

Expected Loss Values for Obtaining a
Response for $L_p = 183$, $L_o = 100$,
and Various Data Base Sizes

Size of Data Base	Expected Loss Values ^a
2	57.3
10	41.8
50	37.3
100	36.6
200	36.3
500	36.1
1000	36.1
2000	36.0
∞	36.0

^aValues expressed to the nearest tenth of
a unit.

TABLE 6
Adjusted Critical Conditional Difficulty
Values for Various Data Base Sizes

Data Base Size	Adjusted Critical Conditional Difficulty Values ^a				
	$N_D = 2$	$N_D = 10$	$N_D = 100$	$N_D = 500$	$N_D \rightarrow \infty$
2	.800	.757	.737	.734	.734
10	.830	.800	.787	.786	.785
50	.837	.810	.797	.797	.797
100	.838	.812	.800	.799	.798
200	.839	.812	.801	.800	.799
500	.839	.813	.802	.800	.800
∞	.839	.813	.802	.800	.800

Note.--When a critical conditional difficulty value of 0.8 is selected for a data base of size N_D .

^aValues expressed to the nearest thousandths of a unit.

dichotomously scored items, as illustrated in Table 1, is obtained.

2. A critical conditional difficulty value and the corresponding data base size are selected. Using Formula 26, values for the critical loss, L_p and L_o , may be computed. The values of L_p and L_o are then used in the testing procedure to compute the expected loss for each decision, using Formulas 23, 24, and 25.

3. A minimum number of items to be presented to each S prior to any item response predictions for S is specified.

4. Responses from S to the minimum number of items, indicated in Step 3, are obtained. The branching strategy, used in the model, requires the difficulty values of the first items presented and the conditional difficulties of the subsequently presented items to be closest to 0.5.

5. For each unpresented item for which no prediction of the correctness of a response has been made, expected loss values for each possible action are determined using Formulas 23, 24, and 25. If the minimum of the three expected loss values indicates prediction to be the best decision, then the corresponding prediction is noted and the item is removed from further consideration. If the expected loss values for any two or more of the actions are equal to the computed minimum expected loss, a decision not to predict is made.

6. For each unpresented item for which no prediction has been made, that is, items for which the expected loss values were minimal for the action to obtain the examinee's response to the item, the conditional difficulty is determined. The branching strategy is

then employed. The item whose conditional difficulty is closest to 0.5 is then presented. After obtaining a response to the item, the procedure beginning with Step 4 is repeated.

7. As additional items are presented to S, the number of observations in the data base upon which conditional difficulties are computed is reduced. If the number of remaining observations in the data base is less than two, or if the denominator of the conditional difficulty ratios equals zero, the remaining items for which responses have neither been obtained nor predicted, are presented to S for his responses.

8. The model suggested by Kalisch (1974) noted that the maximum difference to be permitted between the actual sampling variance of the item conditional difficulty and the variance of the corresponding beta distribution could be specified. Kalisch's (1974) validation study did not employ a maximum difference. In the models investigated in this study, no maximum difference was specified. It was assumed that the corresponding beta distribution was a sufficiently close approximation of the assumed distribution of sample item conditional difficulty values for an examinee.

Model 2

In Model 1, in each case where a decision to predict the correctness of the examinee's response is made, no reconsideration of the prediction is made, even though additional responses have been obtained from the examinee. If the conditional difficulty of an item for an examinee equals either one or zero, additional information

will yield the same value for the conditional difficulty, r/N , provided there are sufficient observations to provide a non-zero denominator. For example, suppose that the conditional difficulty of a particular item equals $10/10$. Also suppose that after obtaining the response to another item with a conditional difficulty of 0.5, it is found that the number of subjects in the data base, with an identical response pattern as the examinee's, equals five. Since all five subjects in the data base answered the former item correctly, the conditional difficulty equals $5/5$, or one. A similar situation exists for items with a conditional difficulty of zero. Item conditional difficulties between zero and one, may increase, decrease, or remain the same, as additional information is obtained. Model 2 is identical to Model 1 with the exception that the former involves recalculating the conditional difficulties and expected losses for items that have not attained values of zero or one, whenever the examinee supplies a response to an item.

The first four steps in the procedure specifying Model 1 also apply to Model 2. The remaining steps in Model 2 are as follows:

5. For each unpresented item for which a previous conditional difficulty value, equaling neither zero nor one, exists, expected loss values for each possible action are determined, using Formula 10. If the minimum of the three expected loss values indicates prediction to be the best decision, and the conditional difficulty of the item equals zero or one, the corresponding prediction is noted and the item is removed from further consideration. If the expected loss values for any two or more of the actions are equal to

the computed minimum expected loss, a decision not to predict is made.

6. For each unrepresented item for which the best decision was to obtain the examinee's response, the conditional difficulty is determined. Of these items, the one with a conditional difficulty closest to 0.5 is presented. After obtaining a response to the item, the procedure beginning with Step 5 is repeated.

7. If the number of remaining observations in the data base is less than two, or if the denominator of the conditional difficulty ratios equals zero, a terminating procedure is initiated. Predictions of the correctness of responses are noted for all items in which the last analysis of expected loss values indicated a prediction to be the best decision. The remaining items are presented to the examinee for his responses.

8. If during the tailored testing procedure, the best decision is a prediction of the correctness of the examinee's response to each of the remaining unrepresented items, the predictions are noted and the testing procedure is terminated.

CHAPTER IV

THE PROBLEM

As previously stated, the purpose of the proposed study was to compare the two tailored testing models, Model 1 and Model 2, and investigate the levels of the variables within the two versions, in terms of actual losses. It was also a purpose of the study to determine if the tailored testing procedures are efficient, that is, whether the total actual losses incurred by employing the models are less than the losses which would be incurred by presenting all the items to each examinee.

Specifically, the following questions concerning the tailored testing models, Model 1 and Model 2, were to be answered by the study:

1. Which of the two versions of the tailored testing models is more efficient?
2. How many items should be presented to the examinee before prediction of the correctness of a response is attempted?
3. What critical conditional difficulty values should be selected by the decision maker?
4. Are the tailored testing procedures using data bases consisting of fewer than 1000 sets of item responses as efficient as the testing procedures using data bases consisting of 2000 sets of

responses?

5. Are the tailored testing procedures using data bases consisting of fewer than 1000 sets of item responses more efficient than presenting all the items to an examinee?

CHAPTER V

METHODS AND PROCEDURES

Computer simulations employing existing data were used for the study. Computer programs were written by the investigator, using FORTRAN IV and COMPASS languages, and were executed on the Control Data Corporation (CDC) 6500 Computer at Florida State University.

The data used in the study were collected in 1973 through the Florida Statewide Assessment Program, which is an objective-referenced testing program (see Footnote 1). One area of testing was ninth grade mathematics, in which three different tests were administered. Each of approximately 15,000 Ss, who were selected by use of matrix sampling procedures, received one of the three tests.

Every item of each test was analyzed by the investigator for readability, correctness, and content validity in relation to the specified corresponding objectives. No two tests contained items corresponding to the same objective, and each objective had more than one item on a test. After removing from consideration ambiguous items and items lacking content validity, nine objectives and the corresponding 33 items of one of the test forms were selected because of the homogeneous content and the inter-relation of the skills required for mastery of the objectives. For the items

selected, 4913 Ss provided responses to at least one of the items. Some examinees apparently did not have a sufficient amount of time to complete answering all the items, as evidenced by the absence of responses to consecutive items at the end of the test. Since some of the items selected appeared as the latter items of the test, all sets of responses in which no response was obtained to two or more of the latter consecutive items were removed from consideration. Responses given by 4844 Ss remained. For the sets of responses remaining, any absence of a response to an item by an examinee was interpreted as an incorrect response.

Using program TESTAT (Veldman, 1967), it was determined that the mean number of correct responses to the 33 items by the 4844 Ss was 24.32; the standard deviation was 9.18; the reliability of the test scores, using KR-20, was .94; and the difficulties of the items ranged from .41 through .93.

The study consisted of two parts. At the beginning of each of the two parts of the experiment, each S's set of item responses was randomly assigned to one of two sets. Throughout each part of the study, Ss' data needed for a data base were selected from one set and the Ss' data to be used for the hypothetical Ss receiving the tailored tests, were selected from the other set.

The remaining portion of this chapter considers the dependent and independent variables and the procedures used in the two parts of the study.

Dependent Variables

Table 3 is a matrix of relative loss values. If for an item, the tailored testing procedure resulted in a prediction that the examinee's response would be incorrect, but a correct response was later obtained from the examinee, the actual loss to the examinee for the item is indicated by L_3 in Table 3, provided the decision was made to predict the student's response to be wrong. The sum of the actual losses for each item of the test constitutes the total actual loss for the examinee.

Three such loss matrices, with actual values specified for the L_i 's in Table 3 were obtained. One matrix was specified by the investigator and the others were obtained independently from two individuals knowledgeable in the area of computer-based instruction and testing. The directions given to the two individuals appear in Appendix C. The dependent variables were the total actual losses per examinee, in relation to each the three evaluation matrices, which are exhibited in Tables 7, 8, and 9 and which are referred to as Evaluation Matrices A, B, and C, respectively.

Some of the questions that were to be answered by the study have been stated using the term "efficiency." Efficiency has been defined in terms of total actual loss. If procedure A yields lower total actual losses than procedure B, with actual losses computed on the basis of the same evaluation loss matrix as procedure A, then procedure A is said to be more efficient than procedure B.

TABLE 7

Evaluation Loss Matrix A

Action	Item Response, if Obtained, Would Be	
	Correct	Wrong
Predict examinee's response to be correct	1	20
Predict examinee's response to be wrong	15	1
Obtain examinee's response to the item	5	5

TABLE 8

Evaluation Loss Matrix B

Action	Item Response, if Obtained, Would Be	
	Correct	Wrong
Predict examinee's response to be correct	1	10
Predict examinee's response to be wrong	5	1
Obtain examinee's response to the item	20	15

TABLE 9
Evaluation Loss Matrix C

Action	Item Response, if Obtained, Would Be	
	Correct	Wrong
Predict examinee's response to be correct	0	3
Predict examinee's response to be wrong	2	0
Obtain examinee's response to the item	1	1

Independent Variables

The four independent variables considered were (1) the models, (2) the critical conditional difficulty values, (3) the number of items to be presented prior to prediction, and (4) the number of Ss for the data base.

The models. As previously mentioned, the tailored testing models, Model 1 and Model 2, were used.

The critical conditional difficulty values. With respect to the assumptions that have been made for this study, the specification of one critical conditional difficulty V , between zero and one, determines the value of the other critical conditional difficulty value, $1-V$. As previously discussed, the critical conditional

difficulty values are dependent upon the size of the data base upon which predictions are being made. In the study being reported, all specifications of critical conditional difficulties assumed the size of the data base to be equal to 10. The critical conditional difficulties selected for the investigation were .80, .85, .90, and .95. Using Formula 26 and letting L_o , the critical loss value for obtaining a response, equal 100, a value for L_p , the critical loss value for predicting the correctness of a response, was determined for each of the selected critical conditional difficulty values. Table 10 shows the critical conditional difficulty values and the corresponding critical loss values for prediction.

TABLE 10
Critical Loss Values for Prediction for
Selected Critical Conditional Difficulties

Critical Conditional Difficulty	Critical Loss Values ^a (L_p)
0.80	209
0.85	358
0.90	673
0.95	1655

Note.--Based upon the critical loss value for obtaining a response as being 100.

^aRounded to the nearest whole number.

With the critical loss values determined, a decision loss matrix such as in Table 2 may be created. Table 11 shows an example of the decision loss matrix for the first case cited in Table 10, when the critical loss value for prediction is 209. From the matrix shown in Table 11, one can construct a graph of the decision loss function, as shown in Figure 4, by setting $L_o = 100$ and $L_p = 209$. In a similar fashion, the decision loss matrices and functions for the critical conditional difficulty values of .85, .90, and .95 may be created.

TABLE 11
Decision Loss Matrix Used for
Critical Conditional Difficulty Level 1

Action	Information Categories	
	$p = 1$	$p = 0$
Predict examinee's response to be correct	0	209
Predict examinee's response to be wrong	209	0
Obtain examinee's response to the item	100	100

Note.--Critical conditional difficulty value of 0.8, based on a data base size of 10.

The levels of the critical conditional difficulties of .80, .85, .90, and .95, referred to as Level 1, Level 2, Level 3, and

Level 4, respectively, were coded for computational purposes as 1, 2, 3, and 4.

Number of items to be presented prior to predictions. In both Model 1 and Model 2, the decision maker must specify how many items will be presented to the examinee prior to any attempt at prediction. The values of concern were limited to the interval [1,5].

Number of subjects for the data base. As previously discussed, the practicality of Model 1 and Model 2 may depend upon the number of S_s from whom responses to all the test items are necessary. The number of S_s to be considered for the data base was restricted to values equal to multiples of 100 between 100 and 2000, inclusive.

Part 1 of the Study

As a means of determining the effects of the versions of the models, the critical conditional difficulty levels, and the number of items presented prior to predictions, a three-factor partially nested univariate analysis of variance was employed for each of the three dependent variables--total actual loss per examinee, as determined from the three evaluation matrices. The mathematical model used was

$$Y_i = \mu + V_j + C_{k(j)} + I_m + VI_{jm} + CI_{km(j)} + \epsilon_{p(jkm)} \quad [28]$$

where Y_i = total actual loss per examinee ($i=1,3$)
 V_j = version of the model ($j=1,2$)
 $C_{k(j)}$ = critical conditional difficulty level ($k=1,4$)
 I_m = number of items presented prior to attempted predictions ($m=1,5$).

The factor, critical conditional difficulties, was nested within versions of the model because of the different roles played by the critical conditional difficulties in the two models. In Model 2, the prediction of the correctness of an item response is conditional, that is, the prediction is reconsidered each time another response is obtained from the examinee. Reconsideration of a prediction does not occur in Model 1. Hence, the decision processes concerning the prediction or obtainment of a response, in relation to the levels of the critical conditional difficulties, are not the same in the two models. Therefore, the mathematical model selected involved a nesting of the critical conditional difficulty factor within the version of the model factor.

For each of the 40 cells a data base consisting of the responses given by 2000 randomly selected Ss to the 33 items was formed. For each of the 40 cells a validation group consisting of 25 randomly selected Ss was also formed.

For each of the six cases determined by the version of the model and the evaluation matrices, the levels of the critical conditional difficulty and the number of items presented prior to prediction that produced the lowest point estimate of total actual loss were determined. The levels were determined by employing regression analysis. On the basis of the three analyses of variance performed, if significant main effects for the conditional difficulty level (C) and the number of item presented to attempted prediction (I) were indicated, but the CI interaction was non-significant, the following

model was used:

$$Y = B_0 + B_1 C + B_2 I + \epsilon. \quad [29]$$

For each regression equation obtained, a lack of fit test and significance of coefficients test were performed ($\alpha = .05$). If it was concluded that the model fit the data and each of the coefficients B_1 and B_2 were significantly different from zero, the model was assumed to adequately represent the data. Regression equations failing to fit the data or failing to have both coefficients significantly different from zero were judged as inadequate. For all cases in which the model given by Formula 29 was not applied and for all cases in which the same model was inadequate, the following model was used:

$$Y = B_0 + B_1 C' + B_2 I' + B_{11} C'^2 + B_{22} I'^2 + B_{12} C'I' + \epsilon, \quad [30]$$

where $C' = C^u$ and $I' = I^v$.

Rather than employing polynomials beyond the second order to provide a more adequate fit of the data to a model and to obtain, if possible, estimates of regression coefficients that were significantly different from zero, a reiterative procedure developed by Box and Tidwell (1962) was used. The procedure provides estimates of power or logarithmic transformations of the independent variables to reduce the residual sums of squares. The model used for this procedure is given by Formula 30. An advantage of this procedure is that the representation of the surface is generally more meaningful when expressed as a second order polynomial equation with transformed independent variables rather than specifying

a higher order model which might involve complex interaction terms. A more detailed discussion of the reiterative procedure appears in Appendix D. Program BMD03R (Dixon, 1971) was used to obtain the estimates of the coefficients and the residual for each observation. These values were used as input to programs written by the investigator to obtain estimates of the power or logarithmic transformations of the independent variables, which were then available as input into BMD03R for determination of another regression equation. The procedure was continued until convergence of the coefficients was indicated or until transformations of the independent variables produced excessively large numbers or numbers very nearly equal to zero.

Using the regression equations based on the linear model, if adequate, and the equations obtained from the reiterative procedure, point estimates of the total actual loss at each of the 20 design points were determined. Within each of the six version by evaluation matrix categories, the levels of the independent variables yielding the minimum total actual loss were identified for use in the second part of the study.

Part 2 of the Study

With the levels of the variables set to values identified in part 1 as producing the most efficient procedures, it was to be determined if equal efficiency would be obtained for smaller data bases. The design used to answer this question was a repeated measures design. Sets of responses were randomly selected. These

sets of responses served as the responses of hypothetical S_s comprising the examinee group. Two sets of responses were independently and randomly selected, one containing 2000 sets of responses and the other containing 100 sets. Using the prescribed levels of the variables, two independent simulations of the tailored testing procedure were conducted for the examinee group, using one data base in one simulation and the other data base in the other simulation. A one-tailed t -test for determination of a difference between two means for correlated samples was used. Let $\bar{L}_{2000}$ and $\bar{L}_{100}$ represent the total actual loss means for the simulations involving the data bases with 2000 and 100 sets of responses, respectively. The null hypothesis was

$$H_0: \bar{L}_{100} - \bar{L}_{2000} = 0$$

with respect to the alternate hypothesis

$$H_1: \bar{L}_{100} - \bar{L}_{2000} > 0.$$

The investigator, applying decision theory to the rejection or non-rejection of H_0 (Hays & Winkler, 1970), was either to accept H_0 or to accept H_1 . In order to use the decision theory approach, careful consideration was given both to Type I and Type II errors.

In the context of the example being discussed, the effect size index, d , for t -tests of means is defined (Cohen, 1969) as

$$d = \frac{\bar{L}_{100} - \bar{L}_{2000}}{\sigma} \quad [31]$$

where σ is the common within-population standard deviation.

Selecting $\alpha = .05$, $\beta = .15$, and $d = .2$, the required number of

paired observations may be determined from power tables (Cohen, 1969, pp. 28-29). Approximately 363 paired observations would have been necessary had the paired observations been uncorrelated. Since the paired observations were to be correlated, an adjusted effect size value, d_1 , is determined from the formula

$$d_1 = \frac{d}{\sqrt{1-r}}, \quad [32]$$

where r is the correlation between the paired observations. Preliminary simulations indicated that the value of r generally exceeded .7. Using the power tables, it was determined that if $\alpha = .05$, $\beta = .15$, $d = .2$, and the number of paired observations to be used equalled 100, d_1 would equal .385 and r would approximately equal .68. Hence, in order to insure a Type II error of at most .15, it was required that the correlation between the 100 paired observations be greater than or equal to .68.

If H_0 was not rejected, it was to be assumed that the tailored testing models employing the data bases of different sizes were equally efficient. If H_0 was rejected in favor of H_1 , it was to be accepted that the procedure employing the smaller data base was less efficient than the procedure using the larger data base. In the case that H_0 was rejected, a simulation involving the 100 sets of responses comprising the examinee group and a randomly selected data base consisting of 200 sets of responses was to be conducted, and a comparison of total actual losses would be made for the 2000-data base and the 200-data base simulations. If the 200-data base failed to provide as efficient a procedure as the 2000-data base,

a data base containing 1000 sets of responses was to be used. If a 1000-data base did not provide as efficient a procedure as the 2000-data base, no other data bases were to be considered for that version by evaluation category. If the 1000-data base provided as efficient a procedure as the 2000-data base, the investigation was to be continued. The smallest data base, which contained an integral multiple of 100 responses, between 200 and 1000, and which provided as efficient a procedure as the 2000-data base, was to be identified.

The final purpose of the study was to determine for each of the six versions of the model by evaluation matrix categories whether a tailored testing procedure employing a data base of fewer than 1000 sets of responses would be more efficient than presenting all the items. In relation to Evaluation Matrices A and C, shown in Tables 7 and 9, the total actual losses for presenting 33 items are 165 and 33, respectively. For Evaluation Matrix B, shown in Table 8, the minimum total actual loss for presenting 33 items occurs if all the items are answered incorrectly. Hence, in relation to Evaluation Matrix B, the minimum total actual loss for presenting all the items is 495.

For each of the six cases, a t-test was conducted to determine the difference between the total actual loss for the tailored testing procedure and the actual loss for presenting all the items. Data bases of size 100 were initially considered. In the same way as for the previous set of t-tests, the smallest data bases of sizes between 100 and 1000, that were more efficient than presenting all

the items were to be identified. Consider the following example. Let $\bar{L}_{100}$ equal the mean total actual loss incurred from a tailored testing procedure, using a base of size 100. Let L represent the total loss incurred by presenting all the items--165, 495, or 33, depending upon the evaluation matrix involved. The null hypothesis tested was

$$H_0: \bar{L}_{100} - L = 0$$

with respect to the alternate hypothesis

$$H_1: \bar{L}_{100} - L < 0.$$

In a practical situation, a computer-based testing program would probably not be initiated unless a moderately large difference between $\bar{L}_{100}$ and L existed. Hence, for the t -tests to be performed for this part of the study, an effect size of $d = .7$ was selected. With a selected value of $\alpha = .01$, the number of observations equal to 100, and $d = .7$, β is less than .005.

CHAPTER VI

RESULTS

Part 1 of the Study

Descriptive statistics for the total actual loss for Model 1 and Model 2 are shown in Tables 12 and 13, respectively. For each of the 40 model by critical conditional difficulty level by items presented prior to attempting prediction categories, three total actual loss means were computed, according to the values in the three evaluation matrices. Hence, the statistics in any one row of Table 12 or Table 13 are based upon the results of the same simulation but reflect the differing evaluation criteria implied by the values in the evaluation matrices. The reader is cautioned that a direct comparison of the statistics obtained by using different evaluation loss matrices may be misleading. A loss determined by using one evaluation matrix cannot be appropriately compared directly to a loss determined by using another evaluation matrix. A comparison of the magnitudes of loss values is only applicable to the loss values computed on the basis of the same evaluation matrix.

Descriptive statistics of the results of the first part of the study are shown in Tables 14 and 15. Included in Tables 14 and 15 are the means, standard deviations, and ranges for the

TABLE 12

Descriptive Statistics of Total Actual Loss for
Model 1 in Part 1 of Study

Minimum Number of Items to be Presented	Evaluation Loss Matrix					
	A		B		C	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Critical Conditional Difficulty--Level 1						
1	140.44	64.60	217.44	58.94	20.04	10.55
2	148.88	74.14	212.08	51.72	21.32	11.96
3	125.84	45.29	213.84	53.92	17.92	7.65
4	148.08	71.58	234.68	58.72	21.68	11.53
5	145.40	74.58	240.20	66.12	21.52	12.32
Critical Conditional Difficulty--Level 2						
1	117.48	40.51	224.48	61.08	17.00	7.21
2	139.68	72.56	253.08	59.08	21.00	11.72
3	92.68	30.33	205.48	67.13	12.84	5.82
4	145.76	64.19	263.76	69.69	22.08	10.70
5	122.08	57.94	233.48	81.19	17.92	10.29
Critical Conditional Difficulty--Level 3						
1	129.84	41.98	259.04	66.71	19.68	7.56
2	131.64	50.67	274.64	93.27	20.32	9.35
3	128.72	50.16	270.92	67.32	19.80	8.91
4	137.20	55.18	274.40	72.89	21.16	9.37
5	134.40	53.00	271.80	69.91	20.56	9.05
Critical Conditional Difficulty--Level 4						
1	132.16	50.27	286.96	55.48	20.63	8.61
2	125.04	60.64	282.44	55.80	19.40	9.81
3	127.88	40.17	300.88	59.93	20.36	7.04
4	129.76	50.93	282.56	67.61	20.20	8.82
5	130.72	37.39	289.72	49.65	20.44	6.41

TABLE 13

Descriptive Statistics of Total Actual Loss for
Model 2 in Part 1 of Study

Minimum Number of Items to be Presented	Evaluation Loss Matrix					
	A		B		C	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Critical Conditional Difficulty--Level 1						
1	124.28	43.38	227.68	71.71	18.08	7.65
2	132.68	57.86	239.88	86.16	19.64	10.28
3	128.72	54.05	240.92	88.04	18.96	9.64
4	132.96	55.20	257.16	85.91	19.96	9.89
5	137.48	61.49	260.48	92.76	20.76	10.90
Critical Conditional Difficulty--Level 2						
1	125.68	44.64	268.68	90.66	19.20	8.38
2	126.72	53.49	249.92	76.42	18.92	9.29
3	129.24	48.63	255.24	82.26	19.40	8.81
4	129.92	49.22	255.72	77.67	19.56	8.81
5	120.72	40.02	276.32	83.97	18.72	7.56
Critical Conditional Difficulty--Level 3						
1	102.36	41.75	243.16	71.13	15.12	7.68
2	101.40	37.51	247.60	81.38	15.12	7.36
3	120.64	31.63	272.64	66.32	18.56	5.98
4	141.20	57.98	296.60	74.28	22.08	9.98
5	134.72	54.69	302.92	87.35	21.40	9.59
Critical Conditional Difficulty--Level 4						
1	136.16	43.76	315.36	62.98	21.96	7.59
2	130.08	45.32	308.68	41.55	20.76	7.44
3	138.32	42.60	306.52	66.60	21.88	7.36
4	128.88	47.51	295.28	61.67	20.40	8.33
5	137.08	41.59	316.68	65.23	22.12	7.23

TABLE 14

Descriptive Statistics of the Results for Model 1 in Part 1 of the Study

Number of Items Before Prediction	Category								
	Items Presented			Correct Predictions			Incorrect Predictions		
	Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range
Critical Conditional Difficulty--Level 1									
1	9.08	3.24	[4,15]	19.84	5.35	[10,28]	4.08	3.38	[0,14]
2	8.80	2.47	[4,14]	19.68	4.98	[7,29]	4.52	3.80	[0,17]
3	9.76	3.05	[4,16]	20.04	4.85	[9,29]	3.24	2.39	[0, 8]
4	10.16	3.38	[5,16]	18.68	5.26	[6,27]	4.20	3.77	[0,18]
5	10.76	3.67	[5,18]	18.40	5.91	[9,28]	3.88	3.76	[0,16]
Critical Conditional Difficulty--Level 2									
1	10.44	3.61	[5,18]	20.08	4.72	[12,28]	2.48	1.90	[0, 8]
2	11.64	3.51	[7,18]	18.08	5.17	[6,26]	3.32	3.76	[0,15]
3	9.52	4.20	[5,24]	22.28	4.69	[9,28]	1.24	1.36	[0, 5]
4	12.40	4.20	[5,20]	16.96	5.75	[9,28]	3.68	3.25	[0,14]
5	10.80	4.75	[5,22]	19.68	6.41	[9,29]	2.56	2.64	[0, 8]
Critical Conditional Difficulty--Level 3									
1	12.12	4.25	[5,21]	18.24	5.09	[8,27]	2.64	1.94	[0, 8]
2	13.68	5.88	[5,27]	16.64	7.06	[4,28]	2.68	2.34	[0, 8]
3	13.40	3.93	[6,22]	17.12	5.57	[6,27]	2.52	2.30	[0, 9]
4	13.48	4.54	[6,25]	16.60	5.69	[2,27]	2.96	2.93	[0,11]
5	12.80	4.20	[6,20]	17.40	5.42	[4,27]	2.84	2.63	[0,11]
Critical Conditional Difficulty--Level 4									
1	13.48	3.59	[10,23]	16.76	4.84	[6,23]	2.40	2.42	[0,10]
2	13.44	3.20	[9,19]	17.44	4.46	[7,23]	2.12	3.18	[0,16]
3	14.96	3.89	[10,23]	16.08	4.78	[7,23]	2.00	2.00	[0, 8]
4	13.72	4.83	[9,27]	16.96	5.64	[5,24]	2.36	2.53	[0,11]
5	13.80	3.16	[9,21]	16.92	3.87	[9,24]	2.32	1.83	[0, 6]

TABLE 15

Descriptive Statistics of the Results for Model 2 in Part 1 of the Study

Number of Items Before Prediction	Category								
	Items Presented			Correct Predictions			Incorrect Predictions		
	Mean	S.D.	Range	Mean	S.D.	Range	Mean	S.D.	Range
Critical Conditional Difficulty--Level 1									
1	10.64	4.94	[4,20]	19.48	5.71	[10,29]	2.88	2.29	[0, 7]
2	11.16	5.35	[4,19]	18.68	6.79	[6,29]	3.16	2.66	[0,11]
3	10.96	5.23	[4,21]	19.12	6.36	[9,29]	2.96	2.54	[0,11]
4	11.84	5.06	[5,21]	18.16	6.46	[4,28]	3.00	2.55	[0, 8]
5	12.12	5.55	[5,24]	17.68	7.10	[5,28]	3.16	2.95	[0,10]
Critical Conditional Difficulty--Level 2									
1	12.92	5.87	[5,26]	17.68	6.54	[5,28]	2.40	2.02	[0, 6]
2	11.60	4.59	[5,20]	18.72	5.86	[9,28]	2.68	2.62	[0, 8]
3	11.96	5.02	[5,25]	18.24	6.11	[3,28]	2.80	2.26	[0, 6]
4	11.92	4.95	[5,21]	18.32	5.89	[9,28]	2.76	2.25	[0, 7]
5	13.60	5.27	[6,26]	17.52	5.70	[7,27]	1.92	1.81	[0, 7]
Critical Conditional Difficulty--Level 3									
1	11.52	4.20	[6,19]	20.16	5.38	[8,27]	1.32	1.85	[0, 7]
2	11.84	4.97	[6,25]	20.00	5.66	[7,27]	1.16	1.41	[0, 5]
3	13.36	4.22	[6,23]	17.64	4.79	[7,27]	2.04	1.48	[0, 5]
4	14.24	4.35	[6,23]	15.80	6.10	[5,27]	3.00	2.86	[0,10]
5	14.88	5.19	[6,27]	15.72	6.10	[4,27]	2.36	2.65	[0,12]
Critical Conditional Difficulty--Level 4									
1	15.92	4.10	[11,25]	14.76	4.95	[5,22]	2.32	2.26	[0, 8]
2	15.40	2.77	[11,21]	15.48	4.00	[8,22]	2.16	2.44	[0,10]
3	14.96	4.60	[9,25]	15.32	5.19	[7,22]	2.68	2.34	[0, 7]
4	14.64	3.92	[9,24]	16.28	5.35	[5,25]	2.12	2.29	[0, 9]
5	16.00	4.41	[10,28]	14.68	4.97	[2,22]	2.36	2.30	[0, 8]

number of items presented to each hypothetical examinee, the number of responses correctly predicted per examinee, and the number of responses incorrectly predicted per examinee, for each of the 40 simulations conducted according to the procedures specified by Model 1 and Model 2.

The results of the three-factor analyses of variance for total actual loss with respect to Evaluation Matrices A, B, and C are shown in Tables 16, 17, and 18, respectively. The levels of the variables were analyzed for each significant effect, using the Newman-Keuls test. With respect to evaluation loss matrix A, a significant conditional difficulty level effect was determined ($F=2.11$, $df=6/960$, $p<.05$). As indicated by Table 19, within each model, greater actual loss was obtained for Level 1 than for at least two of the higher levels. Although a significant conditional difficulty level effect with respect to Evaluation Matrix B was found ($F=19.25$, $df=6/960$, $p<.01$), as shown in Table 17, the conditional difficulty level producing significantly less actual loss was Level 1. Table 20 shows that total actual loss increased as the conditional difficulty level increased, for both models, relative to Evaluation Matrix B.

A model effect was determined for Evaluation Loss Matrix B ($F=14.19$, $df=1/960$, $p<.01$). The mean total actual loss for Model 1 was 254.39, whereas for Model 2 the mean was 271.87. Table 21 indicates that although total actual loss was not significantly increased for any two successive levels of the number of items presented prior to an attempt at prediction, generally, as the

TABLE 16
 Analysis of Variance for Total Actual Loss
 Related to Evaluation Loss Matrix A

Source	df	MS	F
Model (V_j)	1	3463	1.22
Conditional Difficulty Level ($C_{k(j)}$)	6	5987	2.11*
Items (I_m)	4	5238	1.85
VI_{jm}	4	4191	1.48
$CI_{km(j)}$	24	2317	<1
Error ($\epsilon_{p(jkm)}$)	960	2831	

* $p < .05$.

TABLE 17

Analysis of Variance for Total Actual Loss
Related to Evaluation Loss Matrix B

Source	df	MS	F
Model (V_j)	1	74632	14.19*
Conditional Difficulty Level ($C_{k(j)}$)	6	101270	19.25**
Items (I_m)	4	13468	2.56*
VI_{jm}	4	4207	<1
$CI_{km(j)}$	24	886	<1
Error ($\epsilon_{p(jkm)}$)	960	5261	

* $p < .05$.

** $p < .01$.

TABLE 18
 Analysis of Variance for Total Actual Loss
 Related to Evaluation Loss Matrix C

Source	df	MS	F
Model (V_j)	1	6.89	<1
Conditional Difficulty Level ($C_{k(j)}$)	6	175.31	2.07
Items (I_m)	4	173.54	2.05
VI_{jm}	4	111.86	1.32
$CI_{km(j)}$	24	73.93	<1
Error ($\epsilon_{p(jkm)}$)	960	84.50	

TABLE 19

Results of the Newman-Keuls Test for Total Loss for
Critical Conditional Difficulty Levels Within Model
1 and Model 2 Relative to Evaluation Loss Matrix A

Item	Level			
	1	2	3	4
Model 1				
Means	141.73	123.54	132.36	129.11
Non-significantly different means ^a		+	+	+
Model 2				
Means	131.22	126.46	120.06	134.10
Non-significantly different means ^a	+	+	+	+

Note.-- $\alpha=.05$.

^aMeans are non-significantly different if +'s in any one row correspond to those means.

TABLE 20

Results of the Newman-Keuls Test for Total Loss for
Critical Conditional Difficulty Levels Within Model
1 and Model 2 Relative to Evaluation Loss Matrix B

Item	Level			
	1	2	3	4
Model 1				
Means	223.65	236.06	270.16	288.51
Non-significantly different means	(none)			
Model 2				
Means	245.22	261.18	272.58	308.50
Non-significantly different means	(none)			

Note.-- $\alpha=.05$.

TABLE 21

Results of the Newman-Keuls Test for Total
Loss for Number of Items Presented Prior to
Prediction Relative to Evaluation Loss Matrix B

Item	Number of Items				
	1	2	3	4	5
Means	255.35	258.31	258.54	270.02	273.95
Non-significantly different means ^a	+	+	+	+	+

Note.-- $\alpha=.05$.

^aMeans are non-significantly different if +'s in any one row correspond to those means.

number of items was increased, the loss increased. A complete list of total actual loss means and standard deviations for each level of each factor of the analysis of variance model, with respect to each evaluation matrix, is given in Table 35 in Appendix E. Descriptive statistics for each model by item cell appear in Table 36 in Appendix E. The means and standard deviations for total actual loss in each conditional difficulty level by item cell of the analysis have been presented in Tables 12 and 13.

As a means of determining the conditional difficulty level and the number of items that produced the smallest actual loss within each model and with respect to each evaluation matrix, regression analysis was used. The analyses of variance indicated that only with respect to Evaluation Matrix B are both the number of items and the conditional difficulty level effects significant. Also, for Evaluation Matrix B, the item by conditional difficulty level interaction effect was not significant. Therefore a linear regression model, as given by Formula 29, was assumed. For Model 1 the equation obtained was

$$\hat{Y} = 187.84 + 22.87C + 3.19I \quad [33]$$

and for Model 2,

$$\hat{Y} = 201.93 + 20.12C + 6.54I. \quad [34]$$

Tables 22 and 23 are ANOVA tables, indicating lack of fit and significance of coefficients tests for the data and the obtained regression equations within Model 1 and Model 2, respectively. As indicated in Table 23, both coefficients were not equal to zero ($\alpha = .05$) and the assumed linear model did not appear to be

TABLE 22

ANOVA Table for Significance Tests for Coefficients
and for Lack of Fit for Assumed Linear Model
for Evaluation Loss Matrix B and Model 1

Source	df	SS	MS	F
Total	499	2526629		
Regression	2	337082	168540.94	38.26*
B_1	1	326887	326887.00	74.20*
B_2	1	10195	10195.00	2.31
Residual	497	2189547	4405.53	
Lack of Fit	17	79454	4673.76	1.06
Error	480	2110093	4396.03	

Note.--Regression equation--
 $\hat{Y} = 187.84 + 22.87C + 3.19I.$

* $p < .01.$

TABLE 23

ANOVA Table for Significance Tests for Coefficients
and for Lack of Fit for Assumed Linear Model
for Evaluation Loss Matrix B and Model 2

Source	df	SS	MS	F
Total	499	3323728		
Regression	2	295941	147970.50	24.29*
B_1	1	253130	253130.00	41.55*
B_2	1	42811	42811.00	7.03*
Residual	497	3027787	6092.13	
Lack of Fit	17	131855	7756.18	1.29
Error	480	2895932	6033.19	

Note.--Regression equation--
 $\hat{Y} = 201.93 + 20.12C + 6.54I.$

* $p < .01.$

inadequate (Draper & Smith, 1966) since the F -value for lack of fit, 1.29, was not significant at the .05 level. Table 22 indicates that the linear model did not appear to be inadequate, but one of the coefficients was not significantly different from zero. The standard error of the regression coefficient B_2 was 2.10. The 95 percent confidence interval of the coefficient was $[-0.93, 7.31]$. Since it was possible that the coefficient might be either negative or positive, a conclusion as to the best point estimate, that is, the values for C and I that would minimize total actual loss, was not possible. For this case it was decided that the model given by Formula 30 would then be the assumed model.

For the last case discussed and the remaining four cases, an iterative technique presented by Box and Tidwell (1962) was applied as many times as feasible. In some cases convergence of each of the transformation values was evident, but in others convergence did not seem to occur. A more detailed discussion of the procedure appears in Appendix D. In some cases, due to the size of the numbers produced by the transformations, program BMD03R could not be executed again without incurring a computer system error diagnostic, and hence the iterative process was terminated even though convergence was not necessarily evident. Table 24 gives the number of iterations of the transformation process employed, the final transformation values, u and v , and the values of the B_i 's for each of the five cases considered. Tables 25-29 are ANOVA tables indicating lack of fit and significance of coefficients tests for the data and the obtained regression equations.

TABLE 24
Results of Regression After Transformation
of the Independent Variables

Evaluation Loss Matrix	Number of Iterations	Transformation Values		Constant (B_0)	Regression Coefficients			
		u	v		B_1	B_2	B_{11}	B_{12}

Model 1								
A	2	-10.098	0.946	134.27	-7925.81	-4.94	7937.25	1.10
B	3	-1.478	7.688	326.78	-351.34	0.00034	239.33	0.00000
C	1	-3.475	1.369	20.646	-30.474	-0.150	29.875	0.024
								0.087

Model 2								
A	7	3.968	-2.742	137.49	-0.22	-91.22	0.00092	76.14
C	1	2.931	15.758	19.026	-0.105	0.00000	0.0023	0.00000
								0.00000

Note.--Assumed Model-- $Y = B_0 + B_1 C^u + B_2 I^v + B_{11} (C^u)^2 + B_{22} (I^v)^2 + B_{12} C^u I^v + \epsilon$.

TABLE 25

ANOVA Table for Significance Tests for
Coefficients and for Lack of Fit for Regression
Equation Obtained from Iterative Procedure for
Evaluation Loss Matrix A Within Model 1

Source	df	SS	MS	F
Total	499	1631848		
Regression	7	1608650	3223.75	<1
Residual	492	1608650	3269.61	
Lack of Fit	12	53160	4430.00	1.37
Error	480	1555490	3240.60	

Note.--Regression Equation--

$$\hat{Y} = 134.27 - 7925.81C' - 4.94I' + 7937.25C'^{-2} + 1.10I'^{-2} - 0.18C'I', \text{ where } C' = C^{-10.098} \text{ and } I' = I^{0.946}.$$

TABLE 26

ANOVA Table for Significance Tests for
Coefficients and for Lack of Fit for Regression
Equation Obtained from Iterative Procedure for
Evaluation Loss Matrix B Within Model 1

Source	df	SS	MS	F
Total	499	2526629		
Regression	7	359028	51289.7	11.64*
B_1	1	239644	239644	54.39*
B_2	1	4422	4422	1.00
B_{11}	1	97110	97110	22.04*
B_{22}	1	10994	10994	2.50
B_{12}	1	6858	6858	1.56
Residual	492	2167601	4405.69	
Lack of Fit	12	57508	4792.33	1.09
Error	480	2110093	4396.03	

Note.--Regression equation--

$$\hat{Y} = 326.78 - 351.34C' + 0.001I' + 239.38C'^2 + 0.001I'^2 + 0.000C'I', \text{ where } C' = C^{-1.478} \text{ and } I' = I^{7.688}.$$

* $p < .01$.

TABLE 27

ANOVA Table for Significance Tests for
Coefficients and for Lack of Fit for Regression
Equation Obtained from Iterative Procedure for
Evaluation Loss Matrix C Within Model 1

Source	df	SS	MS	F
Total	499	46417.19		
Regression	7	507.09	72.44	<1
Residual	492	45910.10	93.31	
Lack of Fit	12	1577.28	131.44	1.42
Error	480	44332.72	92.36	

Note.--Regression equation--

$$\hat{Y} = 20.646 - 30.474C' - 0.150I' + 29.875C'^2 + 0.024I'^2 + 0.087C'I', \text{ where } C' = C^{-3.475} \text{ and } I' = I^{1.369}.$$

TABLE 28

ANOVA Table for Significance Tests for
Coefficients and for Lack of Fit for Regression
Equation Obtained from Iterative Procedure for
Evaluation Loss Matrix A Within Model 2

Source	df	SS	MS	F
Total	499	1215248		
Regression	7	25764	3680.57	1.51
Residual	492	1189484	2417.65	
Lack of Fit	12	35422	2951.83	1.23
Error	480	1154062	2404.30	

Note.--Regression equation--

$$\hat{Y} = 137.49 - 0.22C' - 91.22I' + 0.00C'^2 + 76.14I'^2 + 0.04C'I', \text{ where } C' = C^{3.968} \text{ and } I' = I^{-2.742}.$$

TABLE 29

ANOVA Table for Significance Tests for
Coefficients and for Lack of Fit for Regression
Equation Obtained from Iterative Procedure for
Evaluation Loss Matrix C Within Model 2

Source	df	SS	MS	F
Total	499	38674.55		
Regression	7	937.52	133.93	1.75
Residual	492	37737.03	76.70	
Lack of Fit	12	1224.31	102.03	1.34
Error	480	36512.72	76.07	

Note.--Regression equation--
 $\hat{Y} = 19.026 - 0.105C' + 0.000I' + 0.002C'^2 +$
 $0.000I'^2 + 0.000C'I'$, where $C' = C^{2.931}$ and
 $I' = I^{15.758}$.

Table 30 indicates the point estimates for each of the item by conditional difficulty level, obtained from each of the resulting regression equations. In Table 30, for each evaluation matrix within a model, the minimal predicted actual loss has been indicated. The conditional difficulty by item level producing the minimal predicted loss for each evaluation matrix was to be identified for use in part 2 of the study. Only for Evaluation Matrix C within Model 2 was the minimal loss value not unique. Because of the extensive amount of time and expense involved in running the computer simulations, conditional difficulty level 3 and item level 1 were selected for Matrix C within Model 2 for use in the second part of the study. By making this selection, one set of simulations could be run and evaluated with respect to both Evaluation Matrices A and C. A similar situation occurred within Model 1. One set of simulations was run for conditional difficulty level 2 and item level 2, but the results were evaluated with respect to both Matrix A and Matrix C.

Part 2 of the Study

Table 31 indicates the descriptive statistics for total actual loss incurred for each evaluation matrix within each model and for each size of data base for which a computer simulation was run. Table 32 shows the correlation between total actual loss incurred with a 2000-data base and a data base of size e for each evaluation matrix within each model. The table also indicates the computed t -values, the probabilities of obtaining the t -values, and the

TABLE 30

Point Estimates of Total Actual Loss for Conditional
Difficulty by Item Levels Within Each Model

Condi- tional Diffi- culty Level	Item Level	Model 1			Model 2		
		Matrix			Matrix		
		A ^a	B ^a	C ^a	A ^a	B ^b	C ^a
1	1	141.68	214.82 ^C	20.01	122.23	228.60 ^C	18.92
1	2	139.91	214.91	20.05	125.34	235.14	18.92
1	3	139.99	216.96	20.25	132.97	241.68	18.92
1	4	141.80	234.36	20.71	135.27	248.23	18.92
1	5	145.26	323.48	21.46	136.17	254.77	18.92
2	1	123.20	231.48	18.03	119.79	248.72	18.36
2	2	121.59 ^C	231.56	17.94 ^C	122.43	255.27	18.36
2	3	121.83	233.26	18.00	130.01	261.81	18.36
2	4	123.79	247.76	18.28	132.29	268.35	18.36
2	5	127.41	321.98	18.84	133.19	274.90	18.36
3	1	130.30	266.79	19.87	113.77 ^C	268.85	17.84 ^C
3	2	128.69	266.86	19.77	114.46	275.39	17.84
3	3	128.93	268.48	19.81	121.80	281.93	17.84
3	4	130.90	282.24	20.08	124.03	288.48	17.84
3	5	134.51	352.70	20.63	124.91	295.02	17.84
4	1	130.41	285.45	20.28	132.83	288.97	20.72
4	2	128.80	285.52	20.18	128.32	295.52	20.72
4	3	129.05	287.10	20.22	135.05	302.06	20.72
4	4	131.01	300.56	20.48	137.11	308.60	20.72
4	5	134.63	369.42	21.03	137.93	315.15	20.72

^aRegression model-- $Y = B_0 + B_1C^u + B_2I^v + B_{11}(C^u)^2 + B_{22}(I^v)^2 + B_{12}C^uI^v + \epsilon$.

^bRegression model-- $Y = B_0 + B_1C + B_2I + \epsilon$.

^cMinimal point estimate for the matrix within the model.

TABLE 31

Descriptive Statistics of Total Actual Loss
Involving Data Bases of Different Sizes

Loss Matrix	Size of Data Base							
	100		200		1000		2000	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Model 1								
A	144.37	68.73	138.39	62.89	134.73	68.35	126.04	56.66
B	219.55	79.92	207.68	59.87	---	---	204.22	57.57
C	20.98	11.30	20.51	10.59	19.89	11.46	18.59	9.75
Model 2								
A	153.52	72.63	146.51	66.16	134.04	45.33	128.72	47.97
B	276.34	103.54	---	---	---	---	270.92	92.91
C	23.72	12.27	22.62	11.19	20.94	8.04	19.87	8.56

TABLE 32

Comparison of Actual Losses for Simulations Employing Data Bases of Size e with Losses Obtained by Using a Data Base of Size 2000

Data Base Size (e)	Correlation	t	p	Conclusion ^{a,b}
Matrix A--Within Model 1				
100	.76	4.08	< .01	Reject H_0
200	.81	3.28	< .01	Reject H_0
1000	.77	1.99	< .05	Reject H_0
Matrix B--Within Model 1				
100	.64	2.48	< .01	Reject H_0
200	.70	0.75	> .05	Fail to reject H_0
Matrix C--Within Model 1				
100	.79	3.38	< .01	Reject H_0
200	.85	3.35	< .01	Reject H_0
1000	.79	1.85	< .05	Reject H_0
Matrix A--Within Model 2				
100	.70	4.75	< .01	Reject H_0
200	.65	3.52	< .01	Reject H_0
1000	.82	1.88	< .05	Reject H_0
Matrix B--Within Model 2				
100	.68	0.68	> .05	Fail to reject H_0
Matrix C--Within Model 2				
100	.76	4.79	< .01	Reject H_0
200	.73	3.56	< .01	Reject H_0
1000	.85	2.34	< .01	Reject H_0

^aNull hypothesis (H_0)-- $\bar{L}_e - \bar{L}_{2000} = 0$
 Alternate hypothesis (H_1)-- $\bar{L}_e - \bar{L}_{2000} > 0$.

^bAlpha level of each test = .05.

conclusions with respect to the null hypothesis

$$H_0: \bar{L}_e - \bar{L}_{2000} = 0$$

and the alternate hypothesis

$$H_1: \bar{L}_e - \bar{L}_{2000} > 0,$$

where $\bar{L}_e$ and $\bar{L}_{2000}$ are the mean actual losses for simulations run with data bases of size e and 2000, respectively. As indicated in Table 32, all null hypotheses tested were rejected except for two cases, both of which involved Matrix B.

The final purpose of the study was to determine for each of the evaluation loss matrices within each model whether a tailored testing procedure employing a data base of fewer than 1000 sets of responses would be more efficient than presenting all the items. The smallest data base considered was of size 100. The hypotheses tested were of the form

$$H_0: \bar{L}_{100} - \bar{L} = 0$$

with respect to an alternate hypothesis of the form

$$H_1: \bar{L}_{100} - \bar{L} < 0,$$

where L was the computed minimum loss for asking all the items. It has previously been mentioned that these computed values for L were 165, 495, and 33, for Evaluation Matrices A, B, and C, respectively. These values may be compared with the actual mean losses incurred for data bases of varying sizes by inspecting the means listed in Table 31. Table 33 provides a tabulation of the computed t -value for each statistical test, and the probability

TABLE 33

Results of One-Tailed t-tests to Compare Total
Actual Loss Obtained for Data Bases of Size
100 and Loss Incurred for Presenting All Items

Evaluation Loss Matrix	t	p
Model 1		
A	- 4.21	< .001
B	- 34.06	< .001
C	- 10.58	< .001
Model 2		
A	- 1.58	> .05
B	- 21.12	< .001
C	- 7.58	< .001

Note.--Null hypothesis-- $\bar{L}_{100} - L = 0$, alternate hypothesis-- $\bar{L}_{100} - L < 0$, where L equals 165, 495, or 33, for matrices A, B, and C respectively.

of obtaining each $\underline{t}$ -value. Since the null hypothesis was not rejected for Matrix A within Model 2, the hypothesis

$$H_0: \bar{L}_{200} - 165 = 0$$

was tested at an α -level of .01 with respect to the alternate hypothesis

$$H_1: \bar{L}_{200} - 165 < 0.$$

The computed $\underline{t}$ -value was -2.79 ($p < .005$). Hence the null hypothesis was rejected.

CHAPTER VII

DISCUSSION

The dependent variables used for every statistical test were total actual losses, which were computed on the basis of the values in the three evaluation matrices. The matrices are shown in Figures 8, 9, and 10. The corresponding values in the matrices differ. Evaluation of a tailored testing procedure using differing evaluation matrices might produce different conclusions concerning the efficiency of the testing procedure. Such was the case in this study.

Evaluation Matrices A and C, shown in Figures 8 and 10, respectively, have different corresponding values, but the values specified in each matrix indicate similar comparative loss patterns. For both matrices the greatest expected loss is specified for the action, predicting an examinee's response to be correct when the item response, if obtained, would have been wrong. For the other possible error in prediction, that is, predicting an examinee's response to be wrong when the item response would have been correct, the expected loss specified in each matrix was two-thirds the expected loss for the former prediction error and greater than any of the other specified loss values within their respective matrices. The smallest expected loss values in each matrix were assigned to the cases

indicating correct prediction. Both matrices indicated that expected losses for obtaining the examinee's response would be greater than expected losses for correct predictions, but less than expected losses for incorrect predictions.

Evaluation Matrix B, shown in Figure 9, indicates that the greatest losses were assigned to the categories corresponding to obtaining an examinee's response. As in Matrices A and C, Matrix B has the lowest expected loss values for cases involving correct prediction. Although in Matrix B the loss values for incorrect prediction are greater than the values for correct prediction, they are lower than losses expected for obtaining an examinee's response.

With respect to all three evaluation matrices, the most efficient tailored testing procedure would be one in which all the item responses would be correctly predicted. With respect to Matrices A and C, a more efficient testing procedure would be one in which all item responses had been obtained rather than one in which all the responses had been incorrectly predicted; whereas, with respect to Matrix B, the opposite is true. The author does not concur with the relative losses indicated in Matrix B. The reader may also fail to concur with the values specified in any one of the matrices. It was not a purpose of the study to determine what the values should be or which of the three evaluation matrices is most appropriate to use, but rather it was intended that the ramifications of employing different levels of the variables considered in the study were to be considered with respect to the differing evaluative criteria.

Which of the Two Tailored Testing Models is More Efficient?

With respect to Evaluation Matrices A and C, both models provided equally efficient testing procedures. Although time was not considered as a variable in the study, the variable may be very important in relation to the testing procedures. It was evident to the investigator that the amount of computer processing time needed to run the computer simulations for Model 2 was much more than that needed for Model 1. Prior to selecting one of the models for implementation, it would appear to be appropriate to consider processing time, cost, and delay of response by the computer to the student for each of the models.

With respect to Evaluation Matrix B, Model 1 was more efficient than Model 2. This result was probably due to the necessity of obtaining more responses to items in Model 2 than in Model 1. Comparing the number of items presented for each item level within each critical conditional difficulty level, the values for Model 1, displayed in Table 14, generally appear to be higher than the corresponding values for Model 2, shown in Table 15. As previously discussed, the expected losses for obtaining responses, specified in Matrix B, were greater than for any other conditions. Hence, it would be expected that if the models were not equally as efficient, Model 1 would be superior.

How Many Items Should be Presented Before Attempting Prediction?

With respect to Evaluation Matrices A and C, presenting one or as many as five items prior to predicting item responses

produced equally efficient testing procedures. Although there was no difference in efficiency, as measured in terms of total loss, for the different number of items presented prior to prediction, the presentation of more than one item prior to prediction might increase the time needed by an examinee to complete a tailored testing session. If a purpose of the tailored testing procedure is to reduce the amount of time needed to test the student, consideration should be given to the effects of the number of items presented initially on total time needed for testing.

Although with respect to Evaluation Matrices A and C no difference in total loss resulted from increasing the number of items presented prior to prediction, with respect to Matrix B, the total loss increased as the number of items initially presented increased. Hence, with respect to Matrix B, the most efficient tailored testing procedures required the presentation of one item prior to prediction.

What Critical Conditional Difficulty Values Should be Selected?

Model 1. With respect to Matrix A, the lowest critical conditional difficulty level (Level 1--critical values of .2 and .8) produced the greatest total loss. The other critical conditional difficulty levels produced equivalent losses.

With respect to Matrix B, total loss increased as the critical conditional difficulty level increased. Hence, the lowest

critical conditional difficulty level produced the most efficient testing procedures.

Unlike the results obtained for Matrices A and B, no critical conditional difficulty level provided smaller total losses than any other level for Matrix C. Although Matrix C was similar to Matrix A in comparative loss patterns, there were greater differences in the magnitudes of pairs of loss values in Matrix A than for most corresponding pairs of loss values in Matrix C.

Model 2. With respect to Matrix A, Conditional Difficulty Level 3 produced minimal total losses. As with Model 1, for Matrix B, total losses increased as the conditional difficulty level increased. Total losses were minimized when $V_1 = .20$ and $V_2 = .80$. For Matrix C, all critical conditional difficulty levels produced equivalent total losses.

Does a Data Base Consisting of Fewer than 1000 Sets of Item Responses Provide as Efficient a Procedure as a Data Base Consisting of 2000 Sets of Responses?

Only with respect to Evaluation Matrix B did data bases consisting of fewer than 1000 sets of item responses provide as efficient a procedure as a data base consisting of 2000 sets of responses. For Model 1, a data base of 200 sets of responses proved to be as efficient as the 2000-data base; for Model 2, a data base as small as 100 sets of responses proved to be as efficient.

Are the Tailored Testing Procedures, Using Data Bases Consisting of Fewer than 1000 Sets of Item Responses, More Efficient than Presenting All the Items to an Examinee?

For both versions of the model and with respect to all evaluation matrices, data bases of sizes less than 1000 sets of item responses provided more efficient testing procedures than presenting all the items of the test to the examinee. Data bases of size 100 provided for superior efficiency of the tailored testing procedures in every case except for Evaluation Matrix A within Model 2. For this one case, a data base of 200 sets of responses was necessary to provide for a tailored testing procedure that was more efficient than presenting all the items to an examinee.

Implications of the Study

The purpose of the study was to determine which of two versions of a tailored testing procedure and what levels of the variables in the testing models produced the most efficient procedure, and what size data base produced a more efficient testing procedure than presenting all the items to an examinee. The conclusions reached differed according to the criteria used to evaluate the testing procedures. Even for similar, although not exactly the same, criteria, conclusions reached were not always the same. Also, the conclusions reached were with respect to one set of test items. Other individuals, using a different set of items, would not necessarily agree to the values used in any of the evaluation matrices employed in the study. Hence, it would appear that simulations similar to those conducted in this

study or studies using human subjects would be appropriate, in order for the decision makers to determine what levels of the variables produce the most efficient procedure.

Although the use of different evaluation criteria resulted in different conclusions, for all three evaluation matrices it was shown that both versions of the tailored testing model could be used with data bases of sizes 100 or 200 to produce testing procedures that were more efficient than presenting all the items to an examinee. Although larger data bases may result in smaller total actual loss, the use of data bases of sizes 100 or 200 may provide a means for tailored testing that is more effective than requiring examinees to answer all the items of a test.

Efficiency of the Testing Procedures

For the two independent variables, critical conditional difficulty and the number of items presented prior to attempting prediction, the effects upon total actual loss were studied using regression. For each evaluation matrix within each model, the levels of the two independent variables producing the smallest point estimate of total actual loss were identified for use in part 2 of the study. The reader is cautioned that these estimates of total loss were not necessarily significantly different from losses incurred using other levels of the independent variables. Hence, although a set of conditions minimizing total loss may have been identified, another set of conditions might produce an equivalent loss.

Identification of the levels of the two independent variables producing the smallest point estimates of total loss was made on the basis of data bases of size 2000. These levels of the independent variables were then used for simulations employing smaller data bases. Future research might determine that the conditions which produce the most efficient testing procedures employing large data bases may also produce the most efficient procedures for smaller data bases. It was determined from the study that in relation to Evaluation Matrices A and C, the testing procedures with identical levels of the independent variables were more efficient when the starting data base was large. However, the results do not rule out the existence of equally efficient testing procedures that use small data bases and levels of the independent variables different from the procedures using larger data bases.

For this study, efficiency was defined in terms of total actual losses, which were determined from the evaluation matrices. The loss values in the matrices were the result of individual judgements, not the result of any data collected for the test items used or the tailored testing procedures employed. Other factors might very appropriately be considered when judging the efficiency of the testing models. For example, a result of the study was that the tailored testing procedures were more efficient than presenting all the items to an examinee. One might wish to measure efficiency by comparing the amount of testing time involved in employing one of the testing models with the time required to present all the items to the examinees. Even though the tailored testing procedures

required, on the average, approximately one-third of the test items to be presented, it should not be assumed that the testing time for a tailored test would be one third the time needed to present all the items. A tailored testing procedure requires computer processing to determine what item should be presented next; whereas an on-line presentation of all the items of a test does not require any elaborate processing to determine the next item to be presented. In a sequential or randomly ordered presentation of items, the next item may reside in a buffer, ready to be displayed as soon as the examinee answers the item presently displayed. There are numerous factors that may affect the computer processing time involved in tailored testing, and each of these factors may therefore affect the efficiency of computer-based tailored testing.

Eliminating the Decision Loss Functions

Formulas 23 and 24 indicate that expected loss for prediction is a function of the conditional difficulty of an item and the critical loss value, which is dependent upon the critical conditional difficulty selected. The expected loss for obtaining a response, as indicated in Formula 25, is also dependent upon the conditional difficulty of an item and the size of the data base upon which the conditional difficulty has been computed. With such dependence of expected loss values on conditional difficulty, one wonders if the decision loss functions might be replaced by decision making based solely upon the conditional difficulty of an item. Table 6 indicates that the adjusted critical conditional difficulty

values change as the size of the data base changes. Even though the decisions are affected by the size of the data base, the range in adjusted critical values may not be sufficiently large to affect the results of the tailored testing procedures. An argument to eliminate the decision loss functions is indeed tenable. But one should also note that the nearly complete dependence of the expected loss values on conditional difficulty is a result of the assumed loss functions--two linear functions and a quadratic function. The testing models do not require the use of these particular functions. Further research may reveal other loss distributions that provide greater efficiency. Hence, it appears that there are two nearly opposite directions for further investigation of the models. The first is to replace the decision loss matrix and corresponding functions with a decision making strategy based directly on conditional difficulty values. The second direction is the investigation of other loss distributions.

Applicability of the Models to Instructional Testing

As previously indicated, the primary intent of the development of the proposed tailored testing models has been for the purpose of providing an efficient computer-based means of testing in instructional situations. Kalisch (1974) indicated the possible applicability of his testing model to determination of mastery of objectives, but no method of assessing mastery was advanced. It appears that with the strong emphasis on objective-referenced testing and mastery learning in present day education, it is

necessary to be able to express the results of the tailored testing procedures in terms of mastery of objectives. Two methods are suggested for possible future research. The first requires no change to the existing models. After employing either testing procedure, each of the responses, either obtained or predicted, for the items corresponding to an objective would be analyzed to determine if the minimum percentage of items would have been answered correctly. The second method would involve prediction of mastery of each objective after each obtained response to an item. With this latter method, conditional probabilities of the mastery of objectives, given the examinee's item responses, would be computed.

It is not the intent of the author to discuss in this paper the numerous practical considerations that must be made prior to implementing a tailored testing model in an instructional situation. The results of the study indicate that the models are feasible for predicting the correctness of item responses, employing relatively small data bases, sizes of 100 or 200. On the basis of the results of this study and Kalisch's (1974) previous study, additional investigations into the refinement and extension of the tailored testing models would appear to be worthwhile.

CHAPTER VIII

SUMMARY AND CONCLUSIONS

Two tailored testing models, intended for use in instructional testing, were developed and investigated. The models specified procedures by which the correctness of an examinee's responses to items of a test are predicted by requiring an examinee to answer as few test items as possible. The study employed computer simulations, using existing data collected from examinees who had answered all the test items.

The models employed item conditional difficulties, beta distributions, loss functions, and decision theory. The models required data bases consisting of S_s ' responses to each item in order to determine conditional difficulties and to compute expected losses for each decision to be made for unrepresented items. These decisions were (1) predict the item response to be correct, (2) predict the response to be wrong, and (3) obtain a response. If the expected loss for a prediction was less than that for obtaining a response, the prediction was noted.

After all predictions were noted, the item with a conditional difficulty closest to 0.5 was presented to the examinee. The procedure continued until each item response was obtained or

predicted. Unlike Model 1, Model 2 required reconsideration of each prediction after each response by an examinee.

The models were compared for efficiency, the total actual loss per examinee. Actual losses were computed from values in an evaluation matrix. Three different independently obtained evaluation matrices were used. Different evaluation matrices applied to the same results of the tailored testing procedures produced different conclusions. In relation to the matrices, the models proved to be equally efficient except in one case in which Model 1 was more efficient than Model 2.

An additional purpose of the study was to investigate the effects of variables in the testing models on efficiency. The presentation of one or as many as five items prior to attempting prediction produced no differential efficiency related to each evaluation matrix, except one in which the presentation of only one item was most efficient.

Another variable investigated was critical conditional difficulty. In order to implement the testing procedures, it would be necessary for the decision makers to preset the critical conditional difficulties. The levels of the critical conditional difficulties considered were .80, .85, .90, and .95. The conclusions were greatly dependent upon the evaluation matrices. With respect to the first matrix, no differential effects occurred. In relation to the second matrix, the lowest level produced the most efficient procedure. With the third matrix, the results were different for the two models. In Model 1 the lowest level

produced the most inefficient procedure; the other levels were equally efficient. In Model 2 the third level provided the most efficient procedure.

The efficiency of the testing procedures employing small data bases of 100 or 200 was equal to the efficiency utilizing large data bases of 2000 for only one matrix. For the other matrices, the smaller data bases reduced efficiency. For each evaluation matrix, the testing procedures employing data base sizes of 100 or 200 were more efficient than obtaining responses to all items from an examinee.

On the basis of the study's results, it was concluded that the model, and the levels of the variables within the models, produce differential efficiency, dependent upon the evaluation matrix. The testing procedures employing small data bases are not necessarily as efficient as the procedures employing large data bases, but for the evaluation matrices used in the study, the tailored testing procedures were more efficient than obtaining an examinee's response to every item.

APPENDICES

APPENDIX A

REDUCTION OF FORMULA 12 to FORMULA 13

APPENDIX A

REDUCTION OF FORMULA 12 TO FORMULA 13

If

$$E(A_i) = \int_0^1 [L_i + (L_j - L_i)x] \frac{(N-1)!}{(r-1)!(N-r-1)!} x^{r-1} (1-x)^{N-r-1} dx, \quad [12]$$

where N and r are integers such that $N \geq 2$ and $0 < r < N$, then

$$E(A_i) = \frac{(N-1)!}{(r-1)!(N-r-1)!} \int_0^1 [L_i (1-x)^{N-r} x^{r-1} + L_j (1-x)^{N-r-1} x^r] dx$$

or

$$E(A_i) = \frac{(N-1)!}{(r-1)!(N-r-1)!} \{ L_i \int_0^1 x^{r-1} (1-x)^{N-r} dx + L_j \int_0^1 x^r (1-x)^{N-r-1} dx \} \quad [12A]$$

Consider the two integrals in Formula 12A. Using integration by parts on the first integral,

$$\begin{aligned} & \int_0^1 x^{r-1} (1-x)^{N-r} dx \\ &= - \frac{x^{r-1} (1-x)^{N-r+1}}{N-r+1} \Big|_0^1 - \int_0^1 - \left[\frac{(1-x)^{N-r+1}}{N-r+1} (r-1)x^{r-2} \right] dx \\ &= 0 + \frac{r-1}{N-r+1} \int_0^1 x^{r-2} (1-x)^{N-r+1} dx. \end{aligned} \quad [12B]$$

The resulting integral in Formula 12B may be reduced by using integration by parts. Repeated use of integration by parts, $r-2$ iterations, produces the expression

$$\frac{(r-1)(r-2)\dots(1)}{(N-r+1)(N-r+2)\dots(N-1)} \int_0^1 x^{r-r} (1-x)^{N-1} dx,$$

which equals

$$\frac{(r-1)!}{(N-r+1)(N-r+2)\dots(N)} \quad [12C]$$

The latter integral in Formula 12A may be simplified by employing integration by parts r times in a similar manner as performed on the first integral in Formula 12. The integration yields the result

$$\int_0^1 [x^r (1-x)^{N-r-1}] dx = \frac{r!}{(N-r)(N-r+1)\dots N} \quad [12D]$$

Substituting the Expressions 12C and 12D into Formula 12A yields

$$E(A_i) = \frac{(N-1)!}{(r-1)!(N-r-1)!} \left\{ L_i \frac{(r-1)!}{(N-r+1)(N-r+2)\dots(N)} + L_j \frac{r!}{(N-r)(N-r+1)\dots(N)} \right\} \quad [12E]$$

Formula 12E may be simplified as follows:

$$E(A_i) = \frac{(N-1)!}{(r-1)!(N-r-1)!} \cdot \frac{(r-1)!}{(N-r+1)(N-r+2)\dots(N)} \left(L_i + \frac{r}{N-r} L_j \right)$$

$$\begin{aligned}
&= \frac{(N-1)!}{(N-r-1)!(N)(N-1)\dots(N-r+1)} \cdot \frac{(N-r)!}{(N-r)!} \left(L_i + \frac{r}{N-r} L_j \right) \\
&= \frac{(N-1)!(N-r)!}{(N-r-1)!N!} \left(L_i + \frac{r}{N-r} L_j \right) \\
&= \frac{1}{N} [(N-r)L_i + rL_j]
\end{aligned}$$

[13]

APPENDIX B

REDUCTION OF FORMULA 16 TO FORMULA 17

APPENDIX B

REDUCTION OF FORMULA 16 TO FORMULA 17

Given the formula

$$E(A_1) = \int_0^1 [4L_1(x-0.5)^2] \left[\frac{(N-1)!}{(r-1)!(N-r-1)!} x^{r-1} (1-x)^{N-r-1} \right] dx \quad [16]$$

where N and r are integers such that $N \geq 2$ and $0 < r < N$, an equivalent form is

$$E(A) = \frac{(N-1)!}{(r-1)!(N-r-1)!} (4L_1) \cdot \int_0^1 (x^2 - x + .25) [x^{r-1} (1-x)^{N-r-1}] dx. \quad [16A]$$

Consider the expression to be integrated in Formula 16A.

$$\int_0^1 (x^2 - x + .25) [x^{r-1} (1-x)^{N-r-1}] dx$$

is equivalent to

$$\begin{aligned} \int_0^1 x^{r+1} (1-x)^{N-r-1} dx - \int_0^1 x^r (1-x)^{N-r-1} dx \\ + \frac{1}{4} \int_0^1 x^{r-1} (1-x)^{N-r-1} dx. \end{aligned} \quad [16B]$$

Each of the three integrals in Formula 16B may be independently

simplified by using integration by parts. The simplification of the second integral

$$\int_0^1 x^r (1-x)^{N-r-1} dx \quad [16C]$$

is explained in Appendix A. The first integral

$$\int_0^1 x^{r+1} (1-x)^{N-r-1} dx \quad [16D]$$

after applying integration by parts, may be expressed as

$$\left| \frac{-x^{r+1} (1-x)^{N-r}}{N-r} \right|_0^1 - \int_0^1 \left[-\frac{(1-x)^{N-r}}{N-r} (r+1)x^r dx \right]. \quad [16E]$$

Formula 16E reduces to

$$\frac{(r+1)}{N-r} \int_0^1 x^r (1-x)^{N-r} dx. \quad [16F]$$

By r successive applications of integration by parts to Expression 16F, the resulting expression may be expressed as

$$\frac{(r+1)(r)(r-1)\dots[r-(r-1)]}{(N-r)(N-r+1)\dots(N-r+r)} \int_0^1 x^{r-r} (1-x)^{N-r+r} dx. \quad [16G]$$

Expression 16G is reducible to

$$\frac{(r+1)!}{(N-r)(N-r+1)\dots(N)(N+1)} \quad [16H]$$

There are $r+1$ factors in the denominator of Expression 16H.

Reversing the order of the factors, it is observed that

$$(N-r)(N-r+1)\dots(N)(N+1) = (N+1)(N)(N-1)\dots(N-r).$$

Hence expression 16H may be expressed as

$$\frac{(r+1)!}{(N+1)(N)(N-1)\dots(N-r)} \quad [16I]$$

Multiplying both the numerator and the denominator of Expression 16I by $(N-r-1)!$ and expressing the results in factorial notation, the simplification of the integral under consideration yields

$$\frac{(r+1)!(N-r-1)!}{(N+1)!} \quad [16J]$$

The third integral in Expression 16B,

$$\frac{1}{4} \int_0^1 x^{r-1} (1-x)^{N-r-1} dx \quad [16K]$$

may be simplified in the same manner as the first integral.

Applying integration by parts to Expression 16K, $r-1$ times, a resulting expression is

$$\frac{(r-1)!}{4(N-r)(N-r+1)\dots(N-1)} \quad [16L]$$

Since the denominator of Expression 16L is equal to $4(N-1)(N-2)\dots(N-r)$, the numerator and denominator of Expression 16L may be multiplied by $(N-r-1)!$ to obtain the expression

$$\frac{(r-1)!(N-r-1)!}{4(N-1)!} \quad [16M]$$

Having obtained simplified expressions for the three integrals of Expression 16B, formula 16A is equivalent to the formula

$$E(A_i) = \frac{(N-1)!}{(r-1)!(N-r-1)!} (4L_i) \left[\frac{(r+1)!(N-r-1)!}{(N+1)!} - \frac{r!(N-r-1)!}{N!} + \frac{1}{4} \frac{(r-1)!(N-r-1)!}{(N-1)!} \right]. \quad [16N]$$

Factoring within Formula 16N yields

$$E(A_i) = \frac{(N-1)!(r-1)!(N-r-1)!}{(r-1)!(N-r-1)!(N-1)!} (4L_i) \left[\frac{(r+1)(r)}{(N+1)(N)} - \frac{r}{N} + \frac{1}{4} \right]. \quad [16O]$$

Formula 16O may be simplified to

$$E(A_i) = L_i \left\{ 4 \frac{r}{N} \left[\frac{r+1}{N+1} - 1 \right] + 1 \right\}. \quad [16P]$$

Combining terms and simplification yield the formula

$$E(A_i) = L_i \left[\frac{4r(r-N)}{N(N+1)} + 1 \right]. \quad [17]$$

APPENDIX C
DIRECTIONS GIVEN TO INDIVIDUALS
SUPPLYING EVALUATION MATRICES

APPENDIX C
DIRECTIONS GIVEN TO INDIVIDUALS
SUPPLYING EVALUATION MATRICES

Tailored testing has been described as a procedure in which an examinee is presented with test items that are appropriate to his level of performance. Such a procedure would be expected to avoid situations in which items that are either "too difficult" or "too easy" are presented to the examinee. By presenting only those items that are appropriate to an individual's performance levels, it would be anticipated that a tailored testing procedure could reduce the amount of time needed for testing examinees. In a computer-based instructional system, in which frequent measurement of students' pre- and post-instructional performances would be conducted, a testing procedure that may provide the necessary information concerning a student's performance level with a reduction in the needed amount of testing time would seem most applicable.

The Problem

Suppose that you are a "decision maker" in an on-going computer-based instructional system. You have previously obtained a substantial amount of data concerning responses by students in the instructional program to the items that appear on pre- and

post-tests. You have decided that a tailored testing procedure will be employed for the on-line tests. The tailored testing procedure requires that the values of certain parameters be set. You, as a "decision maker," are to supply these values.

Parameters in the Tailored Testing Model

Suppose that the complete test now under consideration consists of N items. The tailored testing procedure requires that at least one of the N items be presented to the examinee for a response. For each of the remaining $N-1$ items, the tailored testing procedure provides for three possible actions: (1) predict the examinee's response to the item would be correct, (2) predict the examinee's response to the item would be incorrect, and (3) obtain a response to the item from the examinee. Since the response to an item will be dichotomously scored, the only two possible situations are that the student would either answer the item correctly or incorrectly, had he been presented the item.

The two situations and the three possible actions generate the need for parameters--loss values. Table 1 depicts a loss matrix applicable to the tailored testing model to be employed. The six values (represented by L_i , $i = 1,6$) in the matrix are the expected losses for a specified action, given a particular information category, or situation. The values of the loss matrix may incorporate such factors as: (1) instructional costs, (2) amounts of time for instruction, and (3) effects on the students' morale. In a computer-based testing situation, a greater loss may be

TABLE 1

Matrix of Loss Values
for Determining Actual Loss

Action	Item Response, if Obtained, Would Be	
	Correct	Wrong
Predict examinee's response to be correct	L_1	L_2
Predict examinee's response to be wrong	L_3	L_4
Obtain examinee's response to the item	L_5	L_6

attributed to the increased time and cost needed in presenting an examinee an item and processing his response, than in accurately predicting the correctness of his response. Hence, it would be anticipated that the losses indicated by L_1 and L_4 in Table 1 would be less than L_5 and L_6 . Inaccurate prediction might result in the incorrect placement of a student in an instructional program. In terms of the eventual cost and the inappropriate use of the student's time that might result from incorrect placement and from the ensuing corrective actions, it is anticipated that L_2 and L_3 would be greater than L_1 and L_4 . A specific ordering of the loss values, applicable to all situations, is not possible, since the specification of the values is dependent upon the purpose of the test, the factors considered, and the decision makers.

A Hypothetical Situation

Suppose that the loss values you are to supply are to reflect the situation to be described. An on-line test consisting of 33 items is to be presented to ninth grade students having completed an instructional unit involving objectives measured by the 33 items. The objectives involve homogeneous subject matter in the area of mathematics. Although a hierarchy of skills may be implied by the objectives, a validation of the hierarchy has not necessarily been established. Table 2 indicates the number of items of the test that relate to each objective.

TABLE 2
Number of Items Related to Each Objective

Objective Number	Number of Items
1	6
2	3
3	4
4	4
5	3
6	4
7	4
8	2
9	3

The purpose of the tailored testing procedure to be employed with this test is to present as few items to the student as possible, but accurately predict the correctness of his responses to the remaining items, had the examinee been presented those items.

Assumptions

In specifying the loss values, make the following additional assumptions:

1. Knowing or accurately predicting the correctness of student's responses to the items will be sufficient to determine mastery of the corresponding objectives.

2. Since accurate prediction of the correctness of a student's response to an item is to be considered more efficient than obtaining a response from the student, the loss values supplied for L_1 and L_4 are to be less than the values supplied for L_5 and L_6 .

3. Since inaccurate prediction would be less effective and less desirable than accurate prediction, the values specified for L_2 and L_3 should be greater than L_1 and L_4 .

4. The loss values to be supplied are to represent the losses related to the possible actions that may be taken for each item.

Your Decisions

In relation to the purpose, situation, and assumptions previously discussed, you are asked to supply whole number loss values (0, 1, 2, 3, . . .) for the L_i ($i = 1, 6$) values in Table 1. On the next page is a table similar to Table 1, with blanks substituted for the L_i 's. Please enter the loss values that you, as a "decision maker," would suggest. In a realistic situation, such values would probably be based on previously collected data. But, at the present

time, as a means of determining the feasibility of the tailored testing procedure, you are being asked to use your knowledge, experience, and expertise in computer based testing and instruction to provide what you judge to be appropriate loss values within the hypothetical situation discussed.

Action	Item Response, if Obtained, Would Be	
	Correct	Wrong
Predict examinee's response to be correct	_____	_____
Predict examinee's response to be wrong	_____	_____
Obtain examinee's response to the item	_____	_____

APPENDIX D
A TRANSFORMATION TECHNIQUE OF THE INDEPENDENT
VARIABLES OF A REGRESSION EQUATION

APPENDIX D
A TRANSFORMATION TECHNIQUE OF THE INDEPENDENT
VARIABLES OF A REGRESSION EQUATION

Draper and Smith (1966) have stated that the purpose of transforming the independent variables of a regression model is that the resulting regression model is of a "simple form in the transformed variables, rather than a more complicated one in the original variables" [p. 131]. Unfortunately, the "best" transformations are not generally evident. An iterative technique for deciding upon the transformations of the independent variables has been suggested by Box and Tidwell (1962). The purpose of the technique is to provide a polynomial regression equation of low degree in the transformed variables rather than to contend with polynomials of higher degree in the original variables.

If over a particular region of interest, the function is entirely monotonic, a linear equation with transformed independent variables would be expected to be sufficient. If the function is anticipated to contain an extremum in the region of interest, a polynomial equation of the second degree may be adequate with the use of transformed independent variables.

A regression equation may be expressed in the form

$$Y = f(x_u, \beta) + \epsilon, \quad [35]$$

where the x_u 's are the independent variables and β is the vector containing the constant and the regression coefficients. The error component of the model may be subdivided into error due to lack of fit and pure error. As a means of reducing the error sum of squares due to lack of fit, the reiterative technique of Box and Tidwell may be applied.

The value of the dependent variable may be approximated by the formula

$$Y = f(x_u, \beta) + \sum_{i=1}^k (\alpha_i - 1) \left\{ \partial f(x_u, \beta) / \partial x_{iu} \right\} x_{iu} \ln x_{iu} \quad [36]$$

where k equals the number of independent variables and α_i is a transformational value such that if $\alpha \neq 0$, the transformed value of the independent variable x_u equals $x_u^{\alpha_i}$. If $\alpha_i = 0$, then the transformation is $\ln x_u$. The reiterative technique is employed to obtain the values of α_i .

For the first iteration, a guess is made. If for all i , α_i is assigned a value of one and an iteration of the technique is employed, then new values of the α_i 's may be obtained to produce transformations of the independent variables. The transformed variables may then be used to perform the second iteration.

From Formula 36, if for each iteration the α_i 's are assigned the value of one, the process requires fitting

$$\hat{Y}_u = f(x_u, b). \quad [37]$$

Then for each value of the independent variables,

$$z_{iu} = \left\{ \partial f(x_u, b) / \partial x_{iu} \right\} x_{iu} \ln x_{iu}. \quad [38]$$

On refitting by least squares the x_{iu} 's to the corresponding z_{iu} 's,

$$Y = f(x_u, \beta) + \sum_{i=1}^k (\alpha_i - 1) z_{iu}, \quad [39]$$

estimates of each α_i may be obtained by computing an estimate a_i :

$$a_i = (\sum Y[z_i] / \sum [z_i]^2) + 1, \quad [40]$$

where $[z_i]$'s are the residuals from the equations in x_{iu} 's fitted to the z_{iu} 's.

If for successive iterations the a_i values equal one, convergence has been obtained since no further transformations of the independent variables will occur. After the final transformation, the value of $\hat{Y}$ can be determined at each design point.

A residual sum of squares may be calculated by the formula

$$S_r = \sum (Y - \hat{Y})^2. \quad [41]$$

A sum of squares S_e can be computed by squaring the deviations from each of the corresponding p means, where p equals the number of design points. A difference sum of squares S_f is then calculated by the formula

$$S_f = S_r - S_e. \quad [42]$$

Letting p equal the number of design points, at which there were replications so that there is a total number of n observations, then $n = p + r$. Also, let s equal the number of transformation parameters. Then an approximate test statistic for the adequacy of the fitted transformation function may be given by the expression

$$\frac{r(S_r - S_e)}{(p - f - s)S_e}, \quad [43]$$

which, if the fit is perfect, is approximately distributed as an F distribution with $p-f-s$ and r degrees of freedom. Table 34 shows the degrees of freedom for an analysis of variance after the transformation technique.

TABLE 34
ANOVA Summary for Transformation Technique

Source	Degrees of Freedom
Total	$p + r - 1$
Regression	$f + s - 1$
Lack of Fit	$p - f - s$
Error	r

APPENDIX E
DESCRIPTIVE STATISTICS FOR FACTORS AND INTERACTION
IN THE ANALYSES OF VARIANCE OF PART 1
OF THE STUDY

TABLE 35

Descriptive Statistics for Total Actual Loss Relative to Evaluation Loss Matrices A, B, and C for the Three Analysis of Variance Factors

Level	Evaluation Loss Matrix					
	A		B		C	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
	Model					
1	131.68	57.13	254.59	71.09	19.80	9.64
2	127.96	49.30	271.87	81.54	19.63	8.79
Conditional Difficulty Levels Within Model 1						
1	141.73	67.48	223.65	59.23	20.50	16.03
2	123.54	58.39	236.06	71.15	18.17	14.38
3	132.36	50.49	270.16	74.90	20.30	14.20
4	129.11	48.66	288.51	57.79	20.22	13.68
Conditional Difficulty Levels Within Model 2						
1	131.22	54.92	245.22	86.06	19.48	14.64
2	126.46	47.53	261.17	82.94	19.16	13.57
3	120.06	48.63	272.58	80.27	18.46	13.46
4	134.10	44.37	308.50	60.79	21.42	13.69
Items						
1	126.05	48.26	255.35	75.18	18.97	8.46
2	129.52	59.15	258.54	75.55	19.56	9.95
3	124.00	45.40	258.30	77.57	18.72	8.15
4	136.72	57.41	270.02	74.25	20.89	9.77
5	132.83	54.43	273.95	80.42	20.43	9.45

TABLE 36

Descriptive Statistics for Total Actual Loss Relative
to Evaluation Loss Matrices A, B, and C for the Model by
Item Interaction Term for the Analysis of Variance

Item Level	Model Level			
	1		2	
	Mean	S.D.	Mean	S.D.
Evaluation Loss Matrix A				
1	129.98	50.93	122.12	45.10
2	136.31	65.81	122.72	50.73
3	118.78	44.76	129.23	45.44
4	140.20	61.43	133.24	52.87
5	133.15	57.89	132.50	59.37
Evaluation Loss Matrix B				
1	246.98	66.81	263.72	81.86
2	255.56	72.39	261.52	78.48
3	247.78	73.29	268.83	80.25
4	263.85	69.82	276.19	77.94
5	258.80	71.47	289.10	85.84
Evaluation Loss Matrix C				
1	19.35	8.69	18.59	8.20
2	20.51	10.80	18.61	8.93
3	17.73	8.01	19.71	8.15
4	21.28	10.18	20.50	9.33
5	20.11	9.85	20.75	9.04

REFERENCES

- Birnbaum, A. Some latent trait models. In F. Lord & M. Novick, Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.
- Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-52.
- Box, G. E. P., & Tidwell, P. W. Transformation of the independent variables. Technometrics, 1962, 4, 531-550.
- Cleary, T., Linn, R., & Rock, D. An exploration study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-360. (a)
- Cleary, T., Linn, R., & Rock, D. Reproduction of total test score through the use of sequential programmed tests. Journal of Educational Measurement, 1968, 5, 183-187. (b)
- Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.
- Cronbach, L., & Gleser, G. Psychological tests and personnel decisions. Urbana, Ill.: University of Illinois Press, 1965.
- Dixon, W. J. (Ed.) Biomedical computer programs. Berkley: University of California Press, 1971.
- Draper, N. R., & Smith, H. Applied regression analysis. New York: John Wiley & Sons, 1966.
- Ferguson, R. A model for computer-assisted criterion-referenced measurement. Education, Summer 1970, 91, 25-31.
- Green, B. Comments on tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper & Row, 1970.

- Hansen, D. An investigation of computer-based science testing. In R. C. Atkinson & H. A. Wilson (Eds.), Computer-assisted instruction--A book of readings. New York: Academic Press, 1969.
- Hays, W., & Winkler, R. Statistics: Probability, inference, and decision. Vol. 1. New York: Holt, Rinehart & Winston, 1970.
- Hubbard, J. Programmed testing in the examinations of the National Board of Examiners. (Collected writings) In A. Anastasi (Ed.), Testing problems in perspective. Washington, D. C.: American Council on Education, 1966.
- Kalisch, S. J. A tailored testing model employing the beta distribution and conditional difficulties. Journal of Computer-Based Instruction, 1974, 1, 22-28.
- Linn, R., Rock, D., & Cleary, T. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29, 129-146.
- Linn, R., Rock, D., & Cleary, T. Sequential testing for dichotomous decisions. Educational and Psychological Measurement, 1972, 32, 85-95.
- Lord, F. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.
- Lord, F. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper & Row, 1970.
- Lord, F. Robbins-Monroe procedures for tailored testing. Educational and Psychological Measurement, 1971, 31, 3-31.
- Lord, F., & Novick, M. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.
- Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika, 1969, No. 17.
- Veldman, D. Fortran programming for the behavioral sciences. New York: Holt, Rinehart & Winston, 1967.
- Waters, C., & Bayroff, A. A comparison of computer-simulated conventional and branching tests. Educational and Psychological Measurement, 1971, 31, 125-136.

Wodd, R. Response-contingent testing. Review of Educational Research, 1973, 39, 529-544.

RESUME

Biographical Data

Name: Stanley J. Kalisch, Jr.
Born: March 30, 1943; Yonkers, New York
Marital Status: Married

Home Address: 4251 Parklawn Ave., Apt. 201
Edina, Minnesota 55435
Phone: 612/920-8272

Office Address: Control Data Corporation
ARH215
4201 Lexington Ave., North
St. Paul, Minnesota 55112
Phone: 612/631-0531

Education

1960 - 1962 Manatee Junior College
Bradenton, Florida

1962 - 1963 Florida State University
Tallahassee, Florida
B.S., Mathematics Education

6/65 - 8/65 Florida State University
Tallahassee, Florida

1966 - 1967 Florida Atlantic University
Boca Raton, Florida
M.Ed., Educational Administration and Supervision

1971 - 1974 Florida State University
Tallahassee, Florida
Ph.D., Educational Measurement and Evaluation

Experience

7/1974 - Present Education Consultant, Computer-Based Education
{Course Development Project Manager}
Control Data Corporation
Minneapolis, Minnesota

6/1971 - 6/1974	Florida State University: Doctoral Student, Educational Evaluation and Research Design Program; Research Assistant; Developer and Computer Programmer for computer-based instruction
1969 - 1971	Mathematics Teacher and Curriculum Writer Palm Beach Gardens High School Palm Beach County, Florida
1967, 68, 69 {Summers}	Mathematics Curriculum Writer Palm Beach County, Florida
1964 - 1968	Mathematics Teacher Lake Worth High School Palm Beach County, Florida

Honors

1967	Palm Beach County and District Florida STAR Teacher Award
1969	STAR Teacher Award

Professional Affiliations

American Educational Research Association
Association for the Development of Computer-
Based Instructional Systems
National Council of Measurement in Education
National Council of Teachers of Mathematics

Professional Offices

1966 - 1967	Vice-President Palm Beach County Council of Teachers of Mathematics
1967 - 1969	President Palm Beach County Council of Teachers of Mathematics
1969 - 1971	Vice-President, Secondary Florida Council of Teachers of Mathematics

Professional Committees

- 1970 - 1973 State of Florida
Mathematics Textbook Council
- 4/72 Southern Association of Colleges and Schools
Evaluation Team for South Dade High School
Homestead, Florida
{Chairman: Mathematics Committee}
- 1974 - Present Referee and Reviewer for Journal of Computer-
Based Instruction

Publication

- Kalisch, S. A tailored testing model employing the beta
distribution and conditional difficulties. Journal of
Computer-Based Instruction, 1974 1, 22-28.

Papers Presented at Professional Meetings

- Kalisch, S. Implementation of computer-managed instruction
at Florida State University. {Presented at the meeting
of the Florida Educational Research Association, Tampa,
1973}.
- Kalisch, S., Merrill, P., Towle, N., & Steve, M. The effects
of the availability of objectives and/or rules in a
learner-controlled computer-based task. {Presented at
the meeting of the American Educational Research
Association, New Orleans, 1973}.
- Kalisch, S. A tailored testing model employing the beta
distribution and conditional difficulties. {Presented
at the meeting of the American Educational Research
Association, Chicago, April 1974}.
- Kalisch, S. The use of a modular programming system adaptable
to various instructional needs. {Presented as part of
the symposium, An Approach to Computer Support of
Instruction which is Applicable to a Majority of
Institutions, at the meeting of the American Educational
Research Association, Chicago, April 1974}.
- Kalisch, S. The comparison of two tailored testing models
and the effects of the models' variables on actual loss.
{Presented at the meeting of the American Educational
Research Association, Washington, D.C., April 1975}.

Merrill, P., Kalisch, S., Towle, N., & Steve, M. The interactive effects of the availability of objectives and/or rules on computer-based learning. {Presented at the meeting of the American Psychological Association, Honolulu, 1972}.

Merrill, P., Kalisch, S., & Towle, N. The effects of the availability of prior examples and problems on the learning of rules in a computer-based task. {Presented at the meeting of the American Educational Research Association, New Orleans, 1973}.